POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Analyse des politiques d'affectation d'un service préhospitalier d'urgence par simulation

GABRIEL LAVOIE
Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie industriel

Décembre 2019

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Analyse des politiques d'affectation d'un service préhospitalier d'urgence par simulation

présenté par **Gabriel LAVOIE**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Louis-Martin ROUSSEAU, président
Nadia LAHRICHI, membre et directrice de recherche
Valérie BÉLANGER, membre et codirectrice de recherche
Yannick KERGOSIEN, membre

REMERCIEMENTS

Tout d'abord, je tiens à remercier mes directrices de recherche, Nadia Lahrichi et Valérie Bélanger de m'avoir impliqué dans ce projet et de m'avoir soutenu tout au long de son déroulement. J'ai appris énormément sous votre supervision et vous avez rendu cette expérience des plus agréables.

Ensuite, je dois des remerciements à Luc de Montigny, Karl Rupp-Nantel et Valentina Vulpisi d'Urgences-santé qui m'ont apporté une aide indispensable dans ce projet. Je remercie également les membres du comité directeur de recherche d'Urgences-santé qui ont appuyé ce projet du début à la fin. Sans le soutien d'Urgences-santé, ce projet n'aurait pas été possible.

Je remercie le CRSNG, Polytechnique Montréal et le CIRRELT pour leur soutien financier pendant ma maîtrise.

Enfin, je remercie ma famille et mes amis qui m'ont supporté tout au long de ma maîtrise. En particulier, je souhaite exprimer ma gratitude envers Pierre To et Fannie Jubinville pour leur relecture attentive du mémoire.

RÉSUMÉ

Lorsqu'une urgence médicale survient, le premier réflexe est de composer le numéro d'appel d'urgence. Aussitôt, un processus complexe s'enclenche menant à l'arrivée d'une équipe de techniciens ambulanciers paramédics. Leurs interventions rapides peuvent éviter des drames et sauver des vies. Les services préhospitaliers d'urgence (SPU) sont les organisations derrière ces interventions d'urgence. Leur mission est de fournir des soins préhospitaliers et du transport ambulancier de qualité. La vitesse à laquelle les SPU répondent aux demandes, leurs temps de réponse (TR), est un indicateur important de la qualité de leur service comme l'état de santé des patients peut se dégrader rapidement en l'absence de soins. De plus, même pour les patients non urgents, des temps de réponse courts sont désirables comme les patients souhaitent recevoir de l'aide rapidement.

L'objectif de ce mémoire est de développer un modèle de simulation d'un SPU et de l'utiliser pour développer et pour tester des règles de gestion capables de réduire les temps de réponse. Plus précisément, nous modélisons Urgences-santé, le SPU responsable de Montréal et de Laval, et proposons plusieurs politiques pour améliorer leurs décisions d'affectations. Le modèle est développé en étroite collaboration avec le SPU permettant ainsi de bâtir un modèle réaliste.

L'affectation consiste normalement à choisir l'ambulance à attribuer à chaque patient parmi celles disponibles. Nous proposons d'étendre cette définition en considérant non seulement les ambulances disponibles, mais également celles qui sont indisponibles. En effet, il est possible que ces ambulances occupées soient capables de répondre à un appel plus rapidement que les ambulances libres. Cette approche a été très peu considérée dans la littérature. Nous permettons l'affectation d'ambulances indisponibles dans trois cas : avant le début des quarts de travail, pendant la pause repas et pendant le transfert de patient à l'hôpital. D'après notre modèle, l'inclusion du premier et du troisième groupe dans l'affectation des appels de basses priorités permet de réduire leurs TR de 4.2% et 7.3% respectivement. La deuxième méthode n'améliore pas les temps de réponse que nous expliquons par la redondance entre la position des ambulances qui prennent leur pause repas avec celles des ambulances disponibles.

De plus, nous proposons d'utiliser les ambulances nouvellement disponibles pour améliorer les affectations de basses priorités ce qui mène à une diminution de leurs TR de 7.3%. Au meilleur de notre connaissance, nous sommes les premiers à utiliser cette politique pour des appels de basses priorités. Ces politiques peuvent être combinées, ce qui peut amener des diminutions des TR de basses priorités allant jusqu'à 8.2%. L'effet de ces politiques sur les

appels de hautes priorités est négligeable ou nul. Ces gains sont réalisés par rapport aux règles actuelles d'Urgences-santé dont la politique d'affectation est déjà sophistiquée. Ces résultats nous amènent à recommander l'ajout de ces politiques à leurs politiques. Elles pourraient également être bénéfiques à d'autres SPU.

Notre modèle de simulation a le potentiel d'être appliqué à des politiques de gestion des SPU autres que celles d'affectation. En effet, les politiques de localisation et de relocalisation, de sélection du centre hospitalier, de gestion des effectifs et de spécialisation des ressources pourraient être étudiées à l'aide du modèle.

ABSTRACT

When faced with a medical emergency, our first instinct is to call 911. This puts in motion a complex process that leads to the arrival of one or more teams of emergency responders. Their intervention can save lives and avoid tragedies. Behind the scenes, emergency medical services (EMS) manage these teams. The goal of these organizations is to provide quality emergency prehospital care. The response time (RT), defined as the time required to reach a patient, is a primary quality indicator, because patients' conditions can deteriorate rapidly. Even for non-urgent requests, fast RT is important as patients expect a quick reaction.

The goal of this thesis is to develop a simulation model of an EMS and to use it to measure the impact of new strategies on response time. The model is based on Urgences-santé, the EMS for Montreal and Laval Islands. We propose and evaluate policies to improve their dispatching performance. The model was developed in collaboration with the Urgences-santé, which maximises its fidelity.

Dispatching policies define which ambulance among all available ambulances is chosen to answer a call. We extended this definition to include "near-to-be-available" ambulances as they are sometimes able to reach a call faster than currently available ambulances. This has sparely been done in the literature. Ambulances are considered "near-to-be-available" during three periods: before the beginning of a shift, during the team's lunch breaks, and during transfer of care at the hospital. Using our simulation model, we found that considering the first and third groups in dispatching for low priority calls reduces RT by 4.2% and 7.3% respectively. Considering the second group did not yield improvement to RT, as the locations of ambulances during lunch breaks were too similar to the locations of available ambulances.

In addition, we evaluated a wholly novel strategy to reducing RT based on replacing already-dispatched ambulances with newly-available ambulances to replace ambulances assigned to low priority calls which reduces RT. To be best of our knowledge, this has never been done before. This policy reduced RT for low-priority requests by 7.3%. Combining those policies improved performance further, yielding RT reductions of up to 8.2% for low-priority calls. It is important to note that all evaluated policies had no or negligible impact on the RT for high-priority calls. We recommend the adoption of those policies to Urgences-santé. They might also be applicable to other EMS systems.

The simulation model applications are not limited to the evaluation of dispatching policies. It could be used to evaluate new policies for location and relocation, hospital selection, staff management, and resources specialization.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

CCS	Centre de communication santé
CISSS	Centre Intégré de Santé et de Services Sociaux
CIUSSS	Centre Intégré Universitaire de Santé et de Services Sociaux
ED	Emergency department
EMD	Emergency medical dispatchers
EMS	Emergency medical services
ETA	Estimated time of arrival
ETT	Estimated travel time
FAE	Free-ambulance exploitation
FAELP	Free-ambulance exploitation for low priorities
HP	High priority
LP	Low priority
PABS	Pre-assignment before shift
PADB	Pre-assignment during breaks
PADT	Pre-assignment during transfer
RMU	Répartiteurs médicaux d'urgence
RT	Response time
RTT	Remaining transfer time
SPU	Service préhospitalier d'urgence
TR	Temps de réponse
US	Urgences-santé

CHAPITRE 1 INTRODUCTION

Les services préhospitaliers d'urgence (SPU) assurent un rôle essentiel dans le système de santé. Ils sont responsables de fournir des soins préhospitaliers et du transport par ambulance à la population sur leur territoire. À Montréal et à Laval seulement, plus de 22 000 interventions ambulancières ont lieu chaque mois, dont plus de 19 000 ont mené à des transports. Parmi ceux-ci, plus de 2000 interventions par mois sont en lien avec des arrêts cardio-respiratoires. [1]. L'efficacité des opérations d'Urgence-santé peut avoir des impacts majeurs sur la santé de ses patients. C'est notamment le cas pour les patients en arrêt cardiaque, dont le taux de survie est directement lié à la vitesse d'arrivé des services d'urgences [2]. Dans bien d'autres cas, le temps de réponse (TR), défini comme le temps d'attente avant l'arrivée des paramedics sur les lieux à partir du moment où l'appel est transmis au SPU, a un impact sur la santé des patients [3, 4]. De plus, les patients qui attendent une ambulance peuvent vivre de la détresse et de la douleur, et ce peu importe le niveau urgence d'un point de vue médical [5–7].

En conséquence, réduire les TR est une préoccupation constante pour les gestionnaires des SPU. Le fonctionnement des SPU est centré autour de l'importance de répondre rapides aux demandes de la population. Cependant, malgré cet objectif, la tâche n'est pas simple : il est difficile, voire impossible, de prévoir le lieu et le moment des prochaines demandes. Les SPU doivent donc disposer de politiques de gestion robustes qui assurent la capacité du service à répondre prestement à des demandes qui proviennent de n'importe où sur leur territoire.

Bien que chaque SPU soit géré différemment, il est possible de brosser un portrait global des SPU. Ils sont composés d'une flotte d'ambulances, d'un centre de communication et d'un ou de plusieurs centres opérationnels. Au Québec, les ambulances sont opérées par des techniciens ambulanciers paramédics, alors que le centre de communication emploie des répartiteurs médicaux d'urgence (RMU). Les RMU ont une double tâche : ils doivent répondre aux appels d'urgence et gérer la flotte ambulancière. Les centres opérationnels sont là où les ambulances sont stationnées et entretenues en dehors de leur période d'utilisation. Ces centres servent également de lieux de ravitaillement.

En l'absence d'appels, les ambulances sont placées à des postes d'attentes qui sont généralement répartis sur le territoire couvert par le SPU. Les ambulances ainsi réparties peuvent répondre rapidement aux nouvelles demandes. Les appels d'urgence peuvent être adressés directement au centre d'appel du SPU ou, comme c'est le cas au Québec, être adressés à un centre d'appel responsable de tous les types d'urgence (les centres d'urgence 911). Si c'est

le cas, un répartiteur du centre d'appel général transférera les demandes qui nécessitent une intervention médicale au centre d'appel du SPU. De là, les répartiteurs médicaux d'urgence recueillent les informations nécessaires à l'affectation d'une ambulance. Ils peuvent également donner des instructions à l'appelant sur les premiers soins à apporter en attendant l'arrivée des paramédics. Avec les informations récoltées, une ambulance peut être envoyée sur les lieux de l'incident. Il peut arriver qu'aucune ambulance ne soit disponible immédiatement, et l'appel est alors mis en attente. Lorsqu'une ambulance est affectée à un appel, elle se déplace vers le lieu de l'appel, en utilisant potentiellement une conduite d'urgence accompagnée de sirènes et de gyrophares. Une fois arrivés sur place, les paramédics prodiguent des soins préhospitaliers d'urgence selon les besoins du patient. Par la suite, le patient peut être transporté à un centre hospitalier pour y recevoir des soins supplémentaires. Une fois au centre hospitalier, la responsabilité du patient est transférée au personnel hospitalier et l'équipe de paramédics redevient disponible pour répondre à d'autres appels.

Bien que ce processus soit à priori assez simple, de nombreuses décisions doivent être prises pour le mener à bien. Premièrement, certaines décisions doivent être prises sur le long terme, comme celle de décider du nombre de ressources ambulancières, de paramédics et de RMU. Décider du nombre et de la localisation des centres opérationnels entre également dans les décisions à long terme. À moyen terme se trouvent les décisions quant à la planification des horaires et à la localisation des postes d'attente. Ensuite suivent les décisions à court terme. Ce sont les décisions qui doivent être prises en temps réel par les RMU. La première de ces décisions est l'affectation des ambulances aux appels. Deuxièmement, si un transfert dans un hôpital est requis, un hôpital doit être choisi parmi ceux disponibles. Troisièmement, une fois l'ambulance redevenue disponible, elle peut être envoyée à un des postes d'attente du territoire. Cette décision est désignée sous le nom de « relocalisation ».

La complexité de la gestion d'un SPU combinée à l'importance de sa mission ont mené de nombreux chercheurs du domaine de la recherche opérationnelle à se pencher sur ces services. Plusieurs approches sont possibles par rapport à l'étude de ces systèmes. Le présent mémoire utilise la simulation à événements discrets afin de mesurer l'impact de nouvelles politiques sur la performance d'un SPU. La simulation est communément utilisée dans la littérature [8], car elle permet de modéliser adéquatement les phénomènes aléatoires en lien avec les SPU. Notamment, les caractéristiques de chaque appel, parmi lesquels se trouve la localisation, l'heure de l'appel, la priorité et les besoins médicaux, sont largement imprévisible. Pour pallier à ce problème, nous avons développé un modèle de simulation basé sur Urgences-santé (US). Urgences-santé (US) est le SPU responsable de la région de Montréal et de Laval. Le modèle a été réalisé en collaboration avec US, qui nous a donné accès à l'expertise de ses employés et à ses données afin d'assurer le réalisme du modèle. En plus d'être responsable de

fournir des soins préhospitaliers d'urgence et du transport ambulancier suite à des appels au 911, US est responsable du transport ambulancier inter-hospitalier. Avec une flotte d'environ 165 ambulances et une équipe de plus de 900 paramédics, il répond à environ 375 000 appels d'urgence par année [1]. Il dispose d'un centre de communication santé et de trois centres opérationnels. Par rapport aux autres SPU du Québec, US présente certaines particularités. US est un organisme public responsable d'assurer la coordination des soins préhospitaliers sur son territoire qui gèrent à la fois sa flotte d'ambulance et son centre de communication santé (CCS). Ailleurs au Québec, cette responsabilité relève plutôt des Centres Intégrés Universitaires de Santé et de Services Sociaux (CIUSSS) et des Centres Intégrés de Santé et de Services Sociaux (CISSS) qui font affaire avec des compagnies privées ou des coopératives.

L'objectif du modèle est de mesurer l'impact de changements aux politiques de gestion sur la performance du système, notamment en ce qui concerne les temps de réponse. Présentement, US ne dispose pas d'outil spécifique pour évaluer si de nouvelles politiques auraient des effets positifs sur leur système. Les gestionnaires d'US se basent sur leur jugement afin d'estimer si une nouvelle politique améliorerait la performance de leur opération. Si leur jugement les pousse en ce sens, ils doivent ensuite tester la politique pendant plusieurs semaines afin d'en mesurer les impacts. Avec notre modèle de simulation, il est possible de mesurer les effets d'une modification des politiques par simulation avant même de passer à un essai en situation réelle.

Cette approche présente plusieurs avantages. D'abord, le modèle peut fournir des résultats beaucoup plus rapidement qu'un test en situation réelle, car il faut que les tests réels durent au moins quelques semaines pour présenter des résultats significatifs. Cette rapidité permet également de tester plusieurs alternatives et de les comparer alors que, présentement, une telle démarche demande trop de temps pour être réalisable. Ensuite, la simulation permet d'isoler l'impact du changement, car il est possible de maintenir constantes toutes les variables à l'exception de celles à l'étude. En comparaison, un test dans le système réel peut donner des résultats ambigus, car d'autres changements, tels qu'une hausse des demandes, peuvent se produire pendant la période de test et affecter les résultats.

Malgré ses avantages, la simulation d'un SPU est une tâche complexe. Le fonctionnement des SPU est loin d'être standard. Avant de bâtir un modèle de simulation, il est donc nécessaire d'acquérir une bonne compréhension du système, notamment par le biais de cartographies de processus. Une large quantité de données est également nécessaire pour bien représenter les appels et les ressources disponibles pour y répondre. Ces données doivent être nettoyées et analysées en détail. Advenant la non-disponibilité de certaines données, des hypothèses doivent être formulées pour palier à cela. La modélisation en tant que telle peut également

poser des difficultés, en particulier si le fonctionnement du SPU modélisé est complexe. De plus, la modélisation des déplacements peut également poser des problématiques importantes. Dans notre cas, il était non seulement nécessaire d'estimer les temps de déplacement d'un point à un autre, mais également de connaître la position des ambulances pendant leur déplacement.

Les différents chapitres du mémoire présentent de manière plus détaillée le déroulement du projet de recherche et ses résultats. Le chapitre 2 contient une revue de la littérature scientifique sur les différentes politiques d'affectation des ambulances. Le chapitre 3 décrit la démarche suivie pendant le travail de recherche. Celui-ci suit la démarche standard d'un projet de simulation [9]. Le chapitre 4, présenté sous la forme d'un article, proposes de nouvelles politiques d'affectations des ambulances qui ont été testées à l'aide du simulateur. Le chapitre 5 contient une discussion générale du projet, qui aborde des politiques de gestion autres que celles d'affectation. Pour chacune de ces politiques, nous décrivons leur fonctionnement à US, leur intégration dans le modèle de simulation ainsi que les recherches liées à ces politiques et le potentiel de recherche de notre modèle. Enfin, le chapitre 6 conclut le mémoire.

CHAPITRE 2 REVUE DE LITTÉRATURE

Cette section se consacre à faire une revue de littérature critique au niveau des différentes politique d'affectations. Nous présentons d'abord la politique de l'ambulance libre la plus proche qui est la politique d'affectation la plus courante. L'inclusion d'un système de priorité, un ajout courant à la politique de l'ambulance libre la plus proche, est ensuite abordé. Les politiques reposant sur des informations en temps réel sont ensuite abordées de même que celle utilisant des ambulances occupées. Finalement, des méthodes alternatives à la politique de l'ambulance libre la plus proche sont présentées.

2.1 Politique de l'ambulance libre la plus proche

La politique la plus courante pour les affectations d'ambulances est la *closest-idle policy* ou politique de l'ambulance libre la plus proche. Dans la revue de littérature sur la simulation de SPU d'Aboueljinane [8], sur 24 études, 15 utilisent cette politique. Plus récemment, la revue de littérature de Bélanger *et al.* [10] réaffirme que c'est la politique d'affectation la plus commune. C'est en quelque sorte la politique par défaut pour l'affectation. Les études qui se concentrent sur d'autres aspects de la gestion des SPU utilisent cette politique tandis que celles qui proposent de nouvelles politiques d'affectation s'y comparent.

2.2 Gestion des prioritées

La plupart des modèles présents dans la littérature utilisent un système de priorisation des appels. La priorisation permet de s'assurer que les appels les plus urgents, ceux pour lesquels le temps d'attente est lié à des risques de morbidité ou de mortalité, sont servis le plus rapidement possible. Dans sa forme la plus simple, utiliser un système de priorité affecte uniquement la gestion de la file d'attente des demandes. Cependant, l'utilisation de priorité peut amener plusieurs autres politiques complémentaires.

Un premier ajout à cette politique est l'utilisation de pseudo priorité tel que discuté par Andersson et Varbrand [11]. Cette politique consiste à augmenter la priorité de certains appels après une certaine période d'attente. Par exemple, dans le système avec trois priorités qu'ils étudient, la priorité la moins pressante peut être surpriorisée vers la priorité intermédiaire après un certain temps. Le but de cette politique est d'éviter les temps d'attente excessifs pour les basses priorités, une conséquence possible de l'utilisation de priorité. Il s'agit de la politique utilisée en Suède selon l'article. Cette politique ne fait pas partie du sujet de

l'étude, mais est plutôt un des éléments intégrés dans leur simulateur afin de reproduire le SPU qu'ils étudient.

Une autre option possible est l'utilisation d'un critère de couverture territoriale. Lorsque cette politique est utilisée, l'affectation des appels de basses priorités peut être suspendue si le nombre d'ambulances disponibles sur le territoire tombe sous un certain seuil. Cette politique aide à assurer une couverture du territoire en période d'achalandage élevé et contribue à réduire les temps de réponse pour les appels urgents. Cependant, cela entraîne une hausse des temps de réponse pour les basses priorités. Yoon et Albert [12] étudient cette politique en détails dans le cadre d'un système à deux priorités en considérant plusieurs alternatives au niveau du seuil. Ils testent d'abord cette politique en considérant que les appels de basses priorités qui ne sont pas servis immédiatement sortent du système puis raffinent le modèle en intégrant une file d'attente. Dans les deux cas, ils constatent des gains pour les hautes priorités. Aringhieri *et al.* [13] étudient également cette politique en la comparant à d'autres politiques d'affectation. Ils analysent également l'effet de la politique pour différents seuils. Ils notent que la politique a un effet négatif sur les appels de basses priorités beaucoup plus important que l'effet positif sur les hautes priorités.

Les ambulances assignées à des appels de basses priorités peuvent être réaffectées (*re-routed*) vers des appels de basses priorités. Lim, Mamat et Braunl [14] mesurent l'effet de cette politique et concluent qu'elle permet d'améliorer la performance des appels de hautes priorités significativement au prix d'une baisse de performance pour les appels de basses priorités. Une telle politique est également en place dans le système modélisé par Andersson et Varbrand [11], mais n'est pas étudiée dans leur article. C'est également le cas pour Gendreau *et al.* [15]. Dans leur cas, la politique est uniquement utilisée si l'appel de basse de priorité qui voit son ambulance être réaffectée peut être affecté immédiatement à une ambulance capable de l'atteindre à l'intérieur d'un certain délai. Henderson et Mason [16] modélisent également un système qui utilise cette politique. Ils utilisent le terme *redirect* au lieu de *re-route*. De même, Shin *et al.* [17] modélisent un système qui utilise cette politique. Ils y réfèrent selon l'expression *re-assigned and re-directed*. Le grand nombre de références à cette politique amène à croire qu'elle est assez répandue dans le secteur préhospitalier.

2.3 Décisions dynamiques

Certaines politiques qui peuvent être ajoutées à celle de l'ambulance libre la plus proche exploitent la disponibilité d'information en temps réel sur les ambulances. Ces informations sont disponibles grâce à des systèmes de positionnement à même les ambulances qui communiquent leur position en temps réel au centre de communication santé. Les réaffections

mentionnées dans la section précédente sont un premier exemple de ces politiques comme ils nécessitent de connaître l'emplacement des ambulances pendant qu'ils se déplacent vers un appel.

D'une manière similaire, les ambulances qui se déplacent vers un poste d'attente peuvent être considérées comme disponibles pour des affectations si leur position est connue. Aringhieri *et al.* [13] désignent cette politique sous le nom de *smart assignment* et la comparent à d'autres politiques d'affectation. Ils concluent que cette politique est toujours bénéfique et peut être combinée à d'autres politiques sans problème. C'est, à notre connaissance, la seule étude qui mesure l'impact de cette politique. Cependant, d'autres auteurs ont modélisé des systèmes qui utilisent cette politique dont Gendreau *et al.* [15], Henderson et Mason [16] et Shin [17]. Il est à noter que les systèmes modélisés par ces trois auteurs font également usage de la réaffectation. Les deux politiques sont en effet assez proches l'une de l'autre en ce sens qu'elles reposent sur l'affectation d'ambulances en déplacement.

Une autre politique dans cette catégorie est le *Free-ambulance exploitation* que nous traduisons par le remplacement d'ambulance. Cette politique consiste à remplacer une ambulance déjà en route vers un appel par une ambulance nouvellement disponible qui peut se rendre plus rapidement vers l'appel. C'est une politique dynamique, car elle nécessite de connaître l'emplacement de l'ambulance déjà affecté et d'être capable de comparer son temps d'arrivée estimé à celui de l'ambulance disponible. Cette politique a été proposée par Lim, Mamat et Braunl [14] qui l'étudient conjointement avec la réaffectation. Ils appliquent cette politique uniquement aux appels de hautes priorités sans aborder la possibilité de l'utiliser pour les autres appels. Ils concluent que cette politique permet d'améliorer les temps de réponse des basses priorités et que son utilisation, conjointe avec la réaffectation, les améliore davantage.

2.4 Utilisation d'ambulances occupées

Lee [18] propose le parallélisme *Parallelism*, une politique qui considère les ambulances occupées dans les décisions d'affectation. Il considère que les ambulances qui sont présentement en train de transporter un patient vers un hôpital ou qui transfèrent un patient à l'hôpital peuvent être affectées à un appel. Ces ambulances sont sélectionnées si elles sont en mesure de répondre plus rapidement que les ambulances disponibles dues à leur proximité géographique. Il teste cette politique dans un modèle théorique utilisant une grille de 5 par 5 pour les déplacements. Bien qu'il mentionne que le temps de déplacement est fixe et que le temps à l'hôpital est déterminé par une loi exponentielle, il n'explique pas comment le temps restant à l'affectation est estimé. Son modèle montre que cette pratique peut améliorer les temps de réponse dans de nombreux cas.

Theeuwes [19] modélise un système qui emploie une politique similaire. En effet, le SPU responsable de la région de Brabant-Zuidoost aux Pays-Bas peut utiliser des ambulances dans la phase de transfert du patient à l'hôpital lors d'affectation. Lorsque cette politique est utilisée pour affecter une ambulance à un appel de haute priorité, il est possible de demander aux paramédics d'accélérer le transfert du patient de sorte qu'ils soient disponibles plus rapidement pour répondre à l'appel. Dans le modèle de simulation proposé, lorsque le transfert est accéléré, le temps restant à l'affectation est calculé comme étant la différence entre le temps déjà passé à l'hôpital et 10 minutes. Ce qui peut mener à une fin de transfert instantané si 10 minutes s'étaient déjà écoulées. La valeur de 10 minutes correspond à l'estimation du temps le plus court pour faire un transfert selon des répartiteurs médicaux consultés par l'auteur. Lors de l'affectation, le temps restant au transfert n'est pas considéré. À la place, une pénalité fixe de 7 minutes est ajoutée au temps de réponse des ambulances en centre hospitalier.

2.5 Méthodes alternatives

Plutôt que d'enrichir la politique de l'ambulance libre la plus proche avec de nouvelles politiques, certains chercheurs proposent de rompre complètement avec cette méthode en proposant des alternatives.

Tout d'abord, Gendreau *et al.* [15] proposent d'utiliser la politique de l'ambulance la plus proche uniquement pour les appels de haute priorité. Pour les appels de basse priorité, ils proposent de sélectionner l'ambulance dont l'affectation minimisera les coûts de redéploiement. En effet, ils considèrent, qu'après une affectation, les ambulances libres restantes peuvent être redéployées pour mieux couvrir le territoire. Ces relocalisations étaient utilisées à la suite de 38% des affectations.

Bandara *et al.* [20] proposent d'utiliser une heuristique pour obtenir une politique d'affectation optimale. Leur heuristique vise à maximiser la probabilité de survie. Ils constatent que l'heuristique tend à utiliser l'ambulance la plus proche pour les appels de haute priorité alors que, pour les appels de basse priorité, l'ambulance sélectionnée tend à être la moins occupée, c'est-à-dire celle située dans la zone avec le plus bas volume d'appel.

Nasrollahzadeh [21] utilise un modèle de programmation dynamique dans lequel des appels ne soient pas immédiatement pris en charge même lorsque des ambulances sont disponibles. Cette approche permet de réduire à la fois les temps de réponse et d'augmenter la fraction d'appels répondus à temps. Il observe que les ambulances les plus proches ne sont pas utilisées dans 80% des cas et que les appels ne sont pas affectés immédiatement dans 13% des cas.

Lee [22] a développé un algorithme qui considère les conséquences sur la capacité future du

système à répondre aux demandes lors des décisions d'affectation. Son algorithme repose sur la notion de préparation *Preparedness*. Il mesure cette préparation par zone de service et évalue pour chaque zone à quel point les ambulances disponibles peuvent s'y rendre rapidement. Il oppose cet algorithme à la politique de l'ambulance la plus près qu'il qualifie de gourmand *Greedy* au sens où cette dernière optimise à court terme sans conséquence pour le futur. Après expérimentation, il conclut qu'une approche uniquement basée sur le futur performe moins bien que l'approche gourmande, mais qu'une combinaison des deux améliore la performance du système. Cette combinaison permet de réduire les temps de réponse de 7.9%.

Schmid [23] propose une approche similaire. Pour chaque décision d'affectation, il cherche à minimiser le temps de réponse de l'appel ainsi que les temps de réponse futurs qu'entraîne la décision d'affectation. Il développe un algorithme grâce à des méthodes de programmation dynamique. Par rapport à la politique de l'ambulance la plus proche, il parvient à réduire les temps de réponse de 12.89%.

Jagtenberg *et al.* [24] utilisent, quant à eux, deux méthodes différentes pour améliorer les affectations. La première repose sur un processus de décision markovien pour minimiser le nombre d'appels qui doivent attendre plus d'une certaine période. Comme cette méthode s'applique mal sur des cas plus complexes, ils proposent également une heuristique qui utilise la notion de couverture territoriale pour maximiser le pourcentage d'appels répondus en moins de 720 secondes. Cette approche a cependant des conséquences indésirables comme elles augmentent les temps de réponse pour les appels qui ont dépassé ce seuil.

CHAPITRE 3 DÉMARCHE DE L'ENSEMBLE DU TRAVAIL

Le projet de recherche cherche à répondre à plusieurs objectifs. D'abord, nous souhaitons comprendre le fonctionnement de la gestion de la flotte ambulancière d'US. Ensuite, nous cherchons à reproduire ce système à l'aide d'un modèle de simulation. Enfin, nous visons à concevoir, à tester et à analyser des améliorations aux systèmes mis en place qui peuvent également être applicables dans d'autres SPU. Pour atteindre ces objectifs, les travaux de recherche suivent la démarche standard d'un projet de simulation proposé par Law [9]. Ci-dessous se trouve un résumé des différentes étapes :

1. Formulation du problème et plan d'étude : La première étape du projet a consisté à définir la portée du modèle de simulation et les objectifs du projet. C'est également à cette étape qu'Urgences-santé a donné le feu vert au projet et a débuté son implication.
2. Collecte d'informations et définition du modèle : Afin de bien comprendre le fonctionnement des opérations d'US, nous avons consulté la documentation disponible sur les règles de gestion. De l'observation a aussi été effectuée pendant plusieurs heures dans leur CCS. Également, nous avons posé des questions à un employé responsable de la formation des répartiteurs médicaux d'urgence sur le fonctionnement des opérations. Afin de formaliser notre compréhension de leur processus, nous avons cartographié ceux-ci. C'est également à cette étape que nous avons obtenu des données d'US et défini nos hypothèses.
3. Validation des les hypothèses : Pour cette étape, les cartographies faites dans l'étape précédente ont été présentées à des experts des processus d'US. Ils ont validé la cohérence des cartographies avec leurs processus. Les différentes hypothèses incluses dans le modèle de simulation ont également été discutées avec des employés d'US. Il a été convenu que nos hypothèses étaient raisonnables.
4. Conception et vérification du modèle : Le modèle a été développé dans Arena Simulation 15.1. Le modèle a été conçu par étapes, en partant d'un modèle très simpliste et en ajoutant successivement des composantes afin de le rendre plus réaliste. La vérification et le déverminage du modèle s'est fait au fur et à mesure de son implémentation.
5. Lancement du scénario de validation : Une fois le modèle complété, un premier scénario utilisant les règles normales d'US a été lancé. Ce scénario sert à la fois à valider le modèle et de référence pour les autres scénarios.
6. Validation du modèle : La validation a été divisée en deux étapes. La première a consisté à comparer le scénario de référence aux données utilisées pour construire le

modèle. La deuxième était de faire une démonstration du modèle à des experts des processus d'US. En pratique, l'étape de validation nous a amenés à revenir sur les deux étapes précédentes comme la validation a mis en lumière certaines lacunes du modèle qui ont dû être corrigées.

7. Planification des expériences : Une fois la validation complétée, plusieurs scénarios ont été définis. Initialement, nous avons demandé à des gestionnaires et à des experts des processus d'US de proposer des améliorations à leur système que nous pourrions évaluer. Bien que ces scénarios initiaux aient été complétés et présentés à US, ils ne sont pas présentés dans le présent mémoire comme ils étaient très spécifiques à US et donc d'une pertinence limitée d'un point de vue de recherche. Les scénarios présentés dans la section 4.2 ont été développés dans un deuxième temps et visaient à apporter une contribution scientifique à la gestion des SPU en étudiant de nouvelles politiques d'affectation.
8. Lancement des scénarios : Les scénarios définis à l'étape précédente ont été testés. Nous avons choisi de sauvegarder et d'extraire les données de chaque scénario afin de pouvoir faire des analyses plus poussées en dehors du logiciel de simulation.
9. Analyse des résultats : À partir des données extraites pour chaque scénario, des analyses ont été effectuées à l'aide du langage de programmation R. L'analyse qui en résulte est présentée dans la section 4.4 .
10. Documentation, présentation et utilisation des résultats : En plus de présenter les résultats dans le présent mémoire et dans l'article associé, nous avons également présenté nos résultats à Urgences-santé à plusieurs reprises. Certains résultats préliminaires ont également été présentés lors de congrès Canadian Operational Research Society, Journées de l'Optimisation et Health Care Systems Engineering 2019.

L'article qui constitue le chapitre 4 présente les résultats de cette démarche. Nous y présentons successivement l'étude de cas d'Urgences-santé incluant nos politiques proposées, le modèle de simulation, les résultats de la simulation et une discussion. Comme l'article propose des politiques applicables à d'autres SPU et les testent à l'aide du modèle de simulation, il permet de répondre à nos objectifs de recherche.

CHAPITRE 4 ARTICLE 1: THE VALUE OF "NEAR-TO-BE-AVAILABLE" AMBULANCES IN DISPATCHING POLICIES FOR EMERGENCY MEDICAL SERVICES

Authors : Gabriel Lavoie (Polytechnique Montréal), Valérie Bélanger (HEC Montréal, CIRRELT), Nadia Lahrichi (Polytechnique Montréal, CIRRELT)
Journal : Omega

Abstract : Dispatching policies for emergency medical services (EMS) determine how ambulances are selected and sent to answer emergency calls. They impact the performance of the services since they determine the response time. While the traditional policy consists of sending the "closest-idle ambulance", recent research proposes a broad spectrum of interesting alternatives, including for instance non-available ambulances. These are ambulances already assigned to a request that have not yet reached the location of the emergency call. In this paper, we propose four new policies for dispatching. For the first three, we introduce the notion of "near-to-be-available" ressources (i.e., paramedics are taking their lunch break, paramedics have not yet started their shift, or ambulance is already serving a request and has reached the hospital). One key component of this latter policy is the estimation of the remaining transfer time at the hospital. We propose four different approaches to provide such information. The fourth policy we propose extends the policy of free-ambulance exploitation to low priority requests. We developed a discrete-event simulation model based on the real case of Montreal's EMS *Urgences-santé* to compare these new policies and estimate their impact in practice. Low priority requests represent 19% of all requests in our case study. Our experiments show that all policies (except considering ambulances during lunch break), can lead to a reduction in response time to low priority calls, ranging from 4.2% to 7.3%, with negligeable to no impact response time to high priority calls.

Keywords : Emergency medical services, Dispatching policies, "near-to-be-available" ambulances, Simulation

4.1 Introduction

Emergency medical services (EMS) are responsible for providing emergency pre-hospital care and transport to hospitals in their territories. They ensure the safety of the population in their territory by providing quick emergency response for health related problems. Their services

can save life on a daily basis. The performance of an EMS is often measured by its response time (RT). RT is defined as the time between when the emergency call is received and the arrival of an ambulance to the location of the call. It impacts the health of the patients, including their odds of survival in some cases [2–4]. While there is no universal standard for what constitutes an acceptable RT, several EMS use a target of 8 minutes for their high priority (most urgent) requests [25]. RT targets for low priority (less urgent) requests are less common. A goal of 19 minutes was used in the UK until 2011, while in Singapore an objective of 11 minutes for 80% of calls is used [26, 27].

EMS are composed of a fleet of ambulances and an emergency operations center, where emergency medical dispatchers (EMDs) answer emergency calls and manage the fleet. EMDs make all dispatching decisions, i.e., selecting which ambulance to send to each patient among those available, following the dispatching policies used in-house. During this process, they are assisted by computer-aided software. Determining which dispatching policies are optimal is a complex task and many researches have studied this problem.

The "closest-idle" policy is probably the most widely used dispatching policy both in the literature and in practice [10]. This policy consists of always sending the closest idle ambulance to a request regardless of the priority level. It ensures a quick intervention to the most urgent requests, and is easy to implement. However, recent research has demonstrated that dispatching the closest idle ambulance is not optimal in all cases [11, 15, 22, 23]. This has spurred to the development of new alternatives to support dispatching decisions.

To facilitate the adoption of a policy in practice, some authors propose modifying the closest-idle policy, while still relying on the closeness criteria. Perhaps the most common modification is the use of priority systems [8]. In most studies, we distinguish between two types of requests ; calls requiring an immediate response are given a high priority (HP), while others are given a lower priority. If more than one call is waiting for a dispatch, high priority calls will be served first, by the closest idle ambulance, thus reducing RT for patients requiring the most urgent care. Priority systems can also be used to further develop dispatching policies. In order to prevent excessive waiting time for low priority (LP) calls, Andersson and Värbrand [11] propose increasing the priority of a call after it has been waiting for a given time. They refer to the updated priority as a **pseudo-priority**. It is also possible to **re-route** an ambulance sent to a low priority call to a new higher priority call. As demonstrated in [14], this can reduce RT for HP calls significantly, but requires real-time information about the ambulance location.

Alternatively, Bandara et al. [20] suggests dispatching the closest idle ambulance to HP requests, and the less busy ambulance to LP requests. The goal of this policy is to reduce

RT for HP requests, as well as adress the workload imbalance. Finally, the dispatching of LP calls can be put on hold when the number of ambulances available is lower than a given threshold [13]. This again helps reduce the RT for HP requests, but at the cost of an increase in the RT for LP requests. We refer to this policy as a **coverage-requirement**.

In addition to those policies that rely on priority systems, Aringhieri et al. [13] propose considering ambulances in the redeployment phase (also called relocation), i.e, ambulances that are repositioning and are not assigned to a call. They refer to this policy as a **smart assignment**, and show that it reduces response time for both HP and LP requests. Furthermore, Lim, Mamat, and Braunl [14] propose the **free-ambulance exploitation** (FAE). This policy consists of updating the ambulance assignment based on the closest vehicle : if a new free ambulance is closer to the call, then the previous assignment is canceled. However, the application of this policy is limited to urgent calls. Nevertheless, it improved RT for both HP and LP requests.

Lee [18] proposes a policy called **Parallelism** which considers both available and busy ambulances. The pool of busy ambulances are those that are assigned to a request and are either at a hospital or traveling toward one. The idea is these ambulances may be assigned to the new request if their expected response time is the smallest (i.e., expected remaining time before finishing the request plus the transportation time to the location of the call). This policy is implemented in a theorical model which uses static travel times in a very small transportation network with a single hospital. No precision is provided regarding the method used to calculate the time required to complete the call. Similarly, Theeuwes [19] considers busy ambulances that are assigned to a request but have already reached the hospital. These ambulances may be assigned to the new request if their expected response time is shorter than that of available ambulances by a 7-minute margin. This policy also assumes that the transfer time can be sped up if the busy ambulance is sent to a HP request.

Besides those variations on the closest-idle policy, some researchers have proposed relying on different approaches to improve dispatching. Among all ambulances that can reach a call within a prescribed time frame, Gendreau et al. [15] suggested selecting the one whose dispatch will minimize subsequent relocation costs. Similarly, Andersson and Värbrand [11] propose dispatching the vehicle whose dispatch will lead to the smallest preparedness degradation. In this case “preparedness” is defined as the capacity of the system to answer future demands. Lee [22] also integrates the notion of preparedness to measure the impact of dispatching an ambulance. They show that the policy based on the preparedness measure is able to reduce RT by 7.9% compared to the closest-idle policy. Schmid [23] use dynamic programming to find an optimal dispatching policy. They demonstrate that the use of more flexible dispat-

ching policies contribute to improved RT by 12.9% when compared to the closest idle policy. Jagtenberg et al. [24] formulated the dispatching as a Markov decision problem. They show that the resulting policy can reduce the fraction of late arrival by 37 %, although increasing the average RT. Nasrollahzadeh et al. [21] formulated a dynamic programming model to explore the possibility of dispatching any available ambulance to serve a request at the time it is received, or to queue it until a busy ambulance becomes free. They demonstrate that more than 80% of requests are not served by the closest idle ambulance, and that requests are delayed 13% of the time.

Several studies have shown that the "closest-idle policy" is not always optimal to minimize the response time. Interesting alternatives have been proposed for dispatching. On the one hand, some policies suggest to deviate slightly from the classic closest-idle policy by integrating priorities or considering more ambulances for dispatch. Those policies include, among others, the pseudo-priority, re-routing, and smart assignment, and they have the advantage of being simple enough so as to be easily implemented in practice. On the other hand, other approaches, such as dynamic programming, have been proposed to study dispatching, pointing to the potential benefit of deviating further from the closest-idle policy. However, policies derived from dynamic programming are often difficult to implement in practice [28]. One possible explanation is that policies too distant from the closest-idle policy might not be compatible with the regulations surrounding EMS. Furthermore, EMS dispatch software might not be able to integrate them. A final consideration is that EMS managers might consider that deviating too much from their current approach would pose too many risks.

This paper presents the case study of *Urgences-santé*, one of the largest EMS in Canada. *Urgences-santé* currently uses state-of-the-art dispatching policies which include many of the policies discussed so far. The organization now seeks to find new ways to improve its performance, and is considering modifying current processes. In this paper, we propose four new policies for dispatching. For the first three, we introduce the notion of "near-to-be-available", i.e., paramedics that are taking their lunch break, have not yet started their shift, or are in an ambulance stationed at a hospital during a patient's transfer. One key component of the latter policy is the estimation of the remaining transfer time at the hospital. The fourth policy we propose extends the policy of free-ambulance exploitation to low priority requests. While the concept of "near-to-be-released" shares some similarities with parallelism [18], our proposition goes further by presenting four different dynamic approaches to provide the estimation of the remaining transfer time at the hospital, and studies it in a realistic context. As for other policies we propose, to the best of our knowledge they have not been studied before. We develop a discrete-event simulation model that represents *Urgences-santé's* processes to compare these new policies and estimate their impact in practice.

The remainder of this paper is organized as follows. Section 4.2 presents *Urgences-santé*, their current dispatching policies, and the dispatching policies we propose. Section 4.3 introduces our simulation model and the data. Section 4.4 presents and analyzes results. Section 4.5 provides practical insights regarding the implementation of proposed policies. Section 4.6 provides concluding remarks.

4.2 Case study

4.2.1 EMS overview

Urgences-santé (US) is a public EMS covering the territory of Montréal and Laval, a neighboring suburb. It covers a territory of 744 km^2 and a population of 2.5 million. The EMS is responsible for both emergency calls and patients transportation between hospitals in its territory, which account for about 10% of their activities. While not all EMS manage both tasks, the combination of both has been studied before [16, 29]. US operates a fleet of over 200 vehicles. The ambulances and paramedics are divided between three operational centers in its territory. A notable characteristic of US is its high utilization rate. The utilization rate is the percentage of the time than an ambulance spends assigned to requests during its shift. In the literature, values under 50% are commonly found [17, 30, 31], while for US its values are around 80%.

4.2.2 Current dispatching process

US currently uses a wide range of dispatching policies, which includes many of the policies mentioned in the literature review. US's dispatching processes utilize the concept of closest available ambulance, as well as priorities. In this context, "closest available" is not the same as "closest idle", as it also includes ambulances that are assigned to calls but can be considered for dispatch to higher priorities as defined in the policies they use. A priority system with 9 levels is used. The top 4 priority levels are the most urgent ones and require prompt responses. We refer to those requests as "high-priority" (HP) requests. The bottom five priority levels can wait for over an hour without risk to the patient's health. We refer to these requests as "low-priority" (LP) requests. In this paper, we will only distinguish requests between those two groups. This is to make our results more generalizable, since most priority systems only use two groups. The priority system is used to order the waiting queue. Within the same priority, requests are ordered first come, first served.

Current dispatching processes also include several policies mentioned in the literature review :

— Re-routing : Ambulances traveling to a lower priority request can be used to answer new HP requests. When this happens, the lower priority requests return to the waiting queue. Note that an ambulance can be re-routed from one HP request to another since there are 4 priority levels within this group.
— Smart assignment : Ambulances traveling to a waiting station are considered available for both HP and LP requests.
— Coverage requirement : Ambulances are not dispatched to a LP request unless there are at least 8 ambulances either traveling toward a waiting station or stationed at one.
— Free-ambulance exploitation : When an ambulance becomes available, it can be sent to HP requests already assigned if this would improve the RT by a minimal margin. The exact margin changes by priority.
— Pseudo-priority : Among LP requests, some pseudo-priorities are used, which means that a request with a long waiting time can move in front of requests with a higher priority in the waiting queue.

4.2.3 Proposed dispatching policies

Several dispatching policies proposed in the literature suggest increasing the number of ambulances available for dispatching to improve RT. This is the case for both re-routing and smart assignment. In re-routing, ambulances assigned to LP requests are added to the pool of ambulances available for dispatch, while smart assignment considers ambulances traveling to a station for dispatch. This result is quite intuitive, since adding more ambulances to the pool increases the chance of having one closer to the request. Those policies should lead to increased performance, especially when the utilization rate is high as fewer ambulances are free when a request comes in. Following the same idea, we identify three groups of ambulances that could be added to those considered during dispatch. Including those groups for dispatch leads to three new policies. The novelty of these approaches is that they include groups that are not available to answer to requests immediately, but are expected to be so in the near future, thus introducing the notion of “near-to-be-available”.

The first dispatching policy we propose includes paramedics who have not yet started their shift. For those ambulances, their location and the time of their availability are known in advance. For those ambulances, the estimated time of arrival (ETA) is calculated as the sum of the time remaining before the start of their shift and the estimated travel time (ETT). For other ambulances, the ETA is simply equal to the ETT. The ambulance with the smallest ETA is then sent to serve the request. For ambulances that have not yet started their shift, they are considered to be preassigned to the request.

The second dispatching policy includes paramedics taking their lunch break. Most paramedics teams have a lunch break at some point during their shift. Once they have begun their lunch break (the lunch break can and is regularly cancelled/postponed) they cannot be assigned to any request [1]. We consider that lunch breaks are taken at one of the waiting stations. Once again, the location and the time of the next availability of the team are known in advance. The time remaining for their break is thus added to the ETT to calculate the ETA.

The third policy consists of including ambulances already serving requests that have reached the hospital and are currently stationed at an hospital while the team of paramedics is transferring the patient to the hospital's staff. Those ambulances are considered "near-to-be-available". In this case, the ETA is defined as the sum of the remaining transfer time (RTT) and the ETT. However, the RTT is not known in advance. We propose four different approaches to provide such information.

The three policies described so far propose to pre-assign "near-to-be-available" ambulances if their ETA is the shortest of all ambulances. We refer to those policies as pre-assignment before shifts (PABS), pre-assignment during breaks (PADB) and pre-assignment during transfers (PADT). For all of them, the application is limited to LP requests. Indeed, it is not advisable to use this policy for HP requests since we cannot guarantee the response time of a pre-assigned ambulance. Furthermore, pre-assignment can only occur if the coverage criteria are satisfied both at the time of the pre-assignment and when the ambulance becomes available. This is to ensure that the policy does not penalize HP requests by reducing the number of ambulances that are available to serve them. In addition, when a pre-assigned ambulance becomes available, the pre-assignment only occurs if the ambulance is not near the end of its shift, due for its lunch break, or if it could be dispatched to a HP request through regular dispatching or using by free-ambulance exploitation policy. Those restrictions limit the impact that pre-assignment could have on the paramedics work conditions and on HP requests. If a non-pre-assigned ambulance becomes available, it can be sent to a call that was pre-assigned if it can reach the patient faster than the pre-assigned ambulance. Finally, an ambulance can only be pre-assigned to one request at a time.

The last policy we propose extends the policy of free-ambulance exploitation for low priorities (FAELP). In this case, newly available ambulances can be sent to a LP request already assigned if this would improve response time by at least 5 minutes. This value is used to limit the occurrence of the application of this policy. US already uses similar criteria for HP requests. The FAELP policy can only be considered if there are no unassigned HP requests

1. If a team of paramedic witnesses an emergency event or if someone in distress reaches them during their break, they will take action.

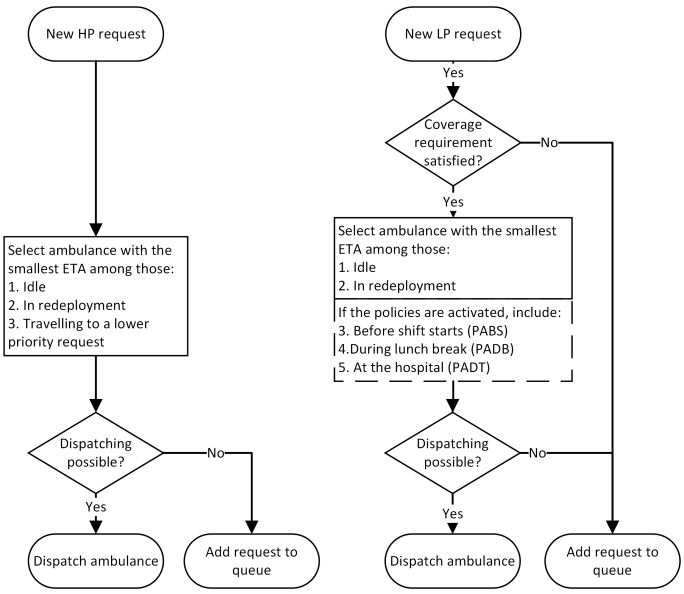

Figure 4.1 Dispatch processes for new HP and LP requests

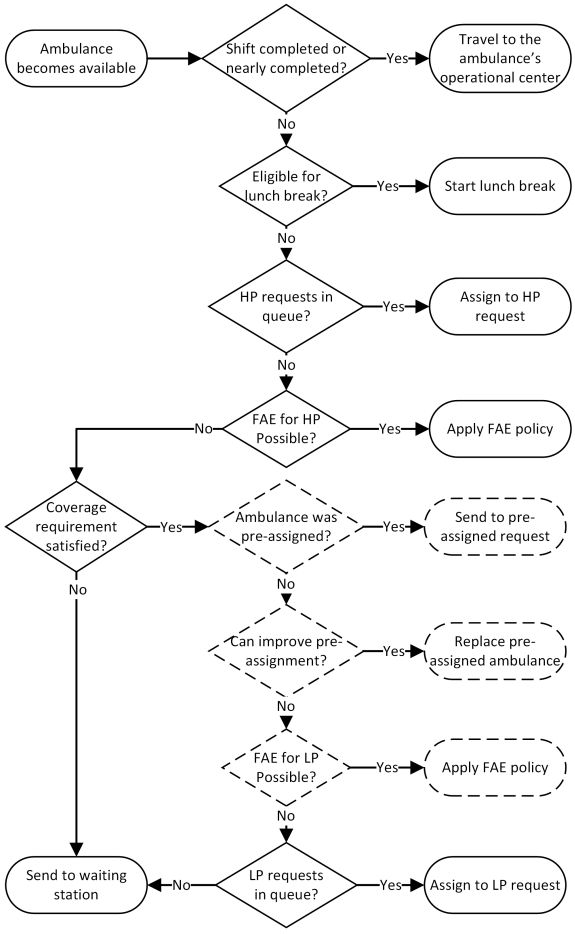

Figure 4.2 Newly available ambulance process

and if free-ambulance exploitation for HP is not possible. When both the FAELP and one of the pre-assigment policies are jointly considered, the pre-assigment policies are applied first.

The processes resulting from the integration of these policies are presented in figures 4.1 and 4.2. The processes and decisions with dotted lines represent new policies. No changes are made to the HP request process, while for LP additional ambulances are considered when one of the pre-assigment policies is activated. When an ambulance becomes available, addititional steps are taken when one of the policies is included. In some cases, one decision event can trigger another one. Following free-ambulance exploitation, the ambulance that was sent first to the call, only to be replaced by an ambulance with a shorter ETA, will, in turn, become available. The use of re-routing will also create an event, as the lower priority request will be once again be unassigned.

4.3 Simulation model

To evaluate the impact of the proposed policies both for high and low priority requests, we developed a discrete-event simulation model and implemented it with Arena Simulation 15.1. We had the full cooperation of US while building the model that reproduces their processes. The EMS took part in every step of the model creation. First, they provided access to their documentation regarding their processes, gave the opportunity to perform observations in their emergency operations center and allowed us to direct questions to an emergency medical dispatcher instructor. This allowed us to form a good understanding of their processes. We produced some process mapping to formalize our understanding, which we validated with a team of experts at US. Following this step, we were provided with a full year of data, which will be detailed in the section below. The model follows the mapped process quite faithfully. When necessary, hypotheses were validated with US to ensure that they were reasonable.

This led to a complex and realistic model capable of simulating a large range of policies including dispatching, relocation, location, hospital selection, and resource management. However, since this paper focuses on dispatching policies, the scope will be limited to this subject.

Conceptually, the model can be represented by figure 4.3. The main module is the core of the simulation model which manages the flow of ambulances and patients. The simulation integrates other modules. Another important module includes all input parameters. Among those, the request history and ambulances schedules will be further discussed in the *Model data* section. The scenario parameters allow us to select which policies to include in a simulation run. Waiting station locations are another important parameter that can impact performance. In our case, we consider that waiting station locations are fixed and known a

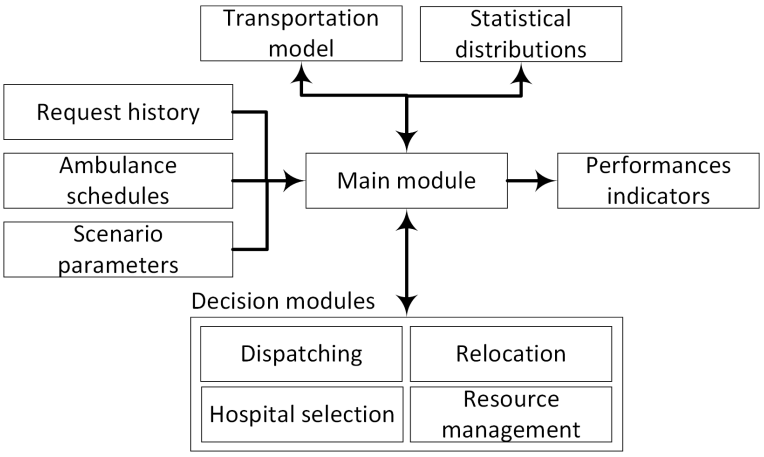

Figure 4.3 Conceptual representation of the simulation model

priori. During the simulation, the main module will call the different decision modules as needed to reproduce the logic that the US medical dispatcher uses when managing the ambulance fleet. In particular, the dispatching module follows the policy described in section 4.2. However, one or more of the proposed policies will be activated according to the selected scenario.Finally, the model has been designed to record different performance indicators. In our case, we will measure the response time for HP requests and the response time for LP requests as they have been identified by US as the key indicator.

4.3.1 Model data

Urgences-santé provided a full year of data to build the simulation model. This data included an anonymized call history and the shift schedule of all ambulances. The call history contained all requests both for emergency calls and interhospital transfers, their location, the time of the call, their priority, which ambulance responded to the call, which hospital they were send to as well as some information required for the hospital selection. Also, tracking times logged into the system for each call were provided. This includes the response time (RT), the travel time to the patient location, the time spent with the patient at the emergency location or at the origin of interhospital transfer, the transport time to the hospital, and the transfer time at the hospital. To ensure patients' anonymity, the GPS coordinates for their pickup location were truncated to a hundredth of a degree. Complete time logs are not available for all requests. For example, a patient might refuse to be transported to a hospital which would cause the transportation time to the hospital and the transfer time to be missing. For the ambulances, we knew at what base they started and finished their work shift, their schedule details and information about their lunch break. In total, the data set included information

about 268,013 requests and 58,260 ambulances shifts.

Since a large amount of data was available, a trace-driven approach was used for both request generation and resource schedules. The requests and ambulances' initial attributes are thus an exact reproduction of those in the data. This approach has been used in other EMS studies using discrete-event simulation [16, 32]. As noted by Aboueljinane [8], the direct use of data ensures the realism of the calls generated. Furthermore, we know that *Urgences-santé* can adjust the number of ambulances deployed according to the intensity of the requests, notably by giving their paramedics the option of working overtime. This means that there is some correlation between the number of calls and the resources. This correlation would be extremely hard to capture with randomly generated requests. We did not, however, use the trace-driven approach for all delays in the system. For the time needed before a newly assigned ambulance could start to travel, and the time spent at the hospital, the data provided was used to fit the probability distribution used to generate those delays.

4.3.2 Transportation model

In general, the transportation model used in an EMS simulation study is one of its most important components. The travel time provided by the model will have a direct impact on the RT and on the time required to complete a request. Inappropriate modeling of transportation times can lead to inaccurate performance measurements. A well-known example is the one of Carson & Batta who predicted an improvement to average RT of 30%, but only achieved a 6% improvement in reality [33]. They attribute the difference between the predicted and the real performance to the transportation model used. In our case, a special constraint is that we need to know the location of the ambulance while they are traveling to any destination in order to use the re-routing, free-ambulance exploitation and smart assignment policies.

Since we were provided patients' locations with a precision of a hundredth of a degree of longitude and latitude, we chose to base the transportation model around those restrictions. A grid was drawn over the territory of *Urgences-santé*, and was divided so that each square of the grid was associated with one of the unique combinations of truncated coordinates. This led to a 31 by 50 grid where 950 squares were part of US territory. Each square measures about 780 by 860 meters. A node is positioned inside of each square on either the main road within the squares or on it's largest intersection. Each node is linked to its neighboring nodes by an edge unless there is a physical obstacle preventing direct transport by car between the two nodes. This creates a graph composed of 950 nodes and 3,351 edges. Figure 4.4 shows a section of the resulting graph ; the nodes separated by the river are only linked if a bridge is present.

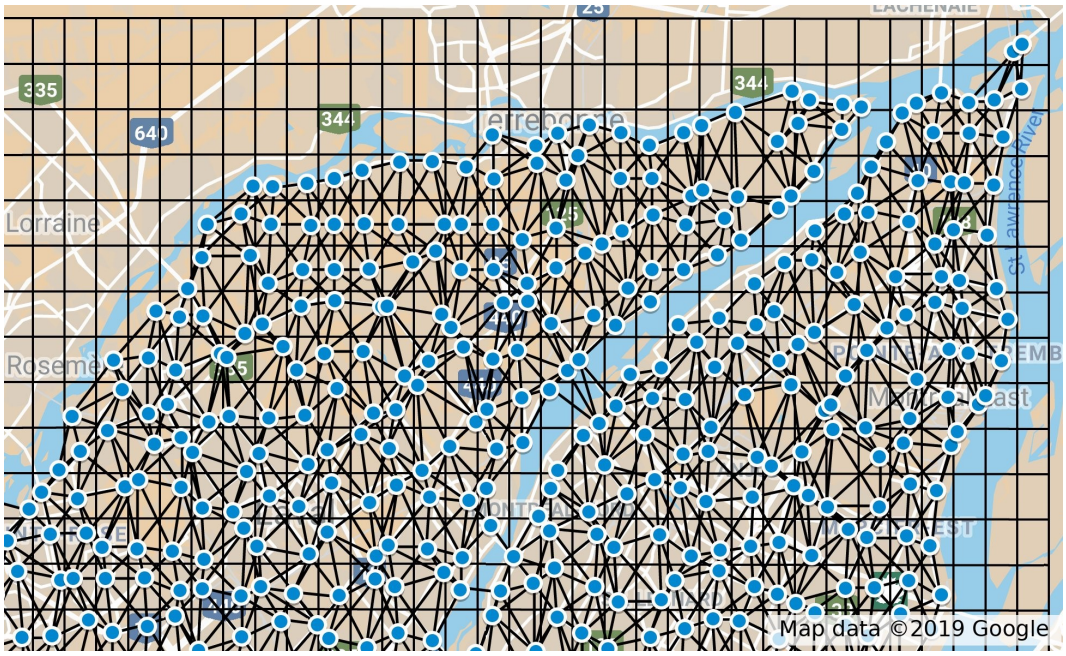

Figure 4.4 A section of the transportation graph

To go from this basic graph to a usable transportation model, a base weight was given to each node using Google's Distance Matrix API. The shortest path for each pair of nodes was then calculated. In the model, ambulances travel along the graph using the nodes, which makes their location known at all times. The base weights found using the API are not used directly. Instead, using the transportation data available, a corrective factor was applied to those travel times based on the priority of the request and the hour of the day. The priority is used to represent the use of lights and sirens. In some cases, ambulances have to travel outside of the territory either to pick up a patient, or to deliver a patient to a hospital. When this happens, the travel time used is the average of the historical travel time to this hospital.

4.3.3 Validation

To validate the simulation model, we looked at both face validity and historical data comparison. For face validity, we made a live demonstration of the model to several *Urgences-santé* executives and process experts. We demonstrated how different processes were handled inside of the simulator with different examples. No significant difference was found between *Urgences-santé* and the simulated processes. To ensure that the model provided similar results to the data, the response times of both were compared for 60 replications of three months, plus a 20-day warm-up period. Figures 4.5 and 4.6 compare the cumulative distribution of response time extracted from the data, and the one obtained using simulation,

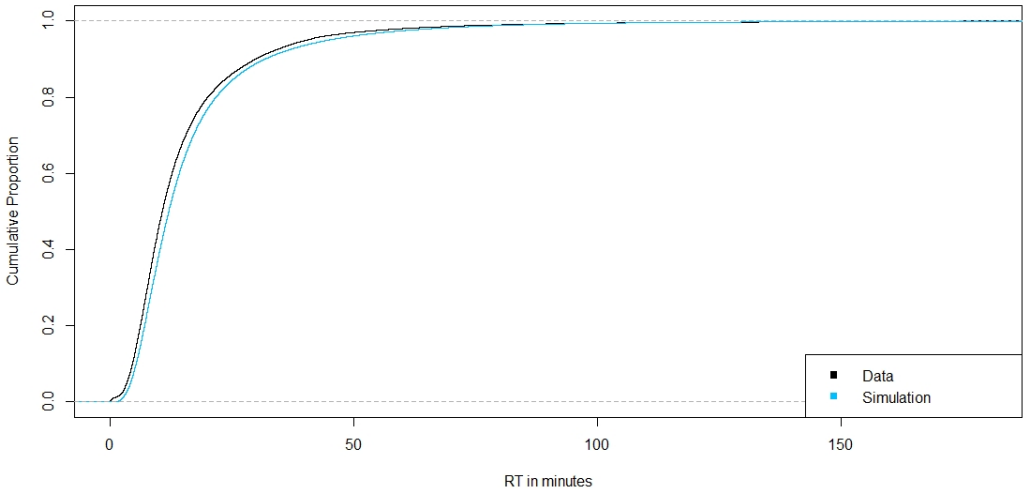

Figure 4.5 Validation of response time for HP requests

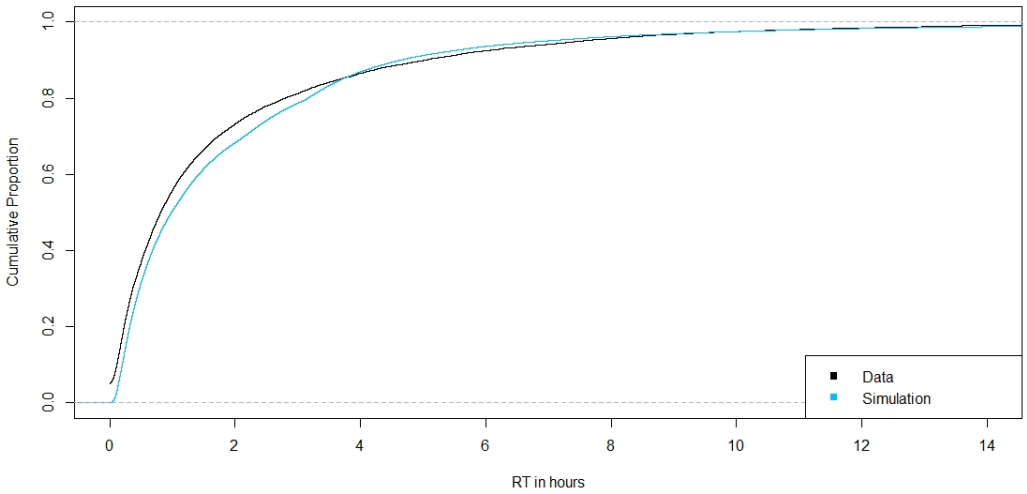

Figure 4.6 Validation of response time for LP requests

for both HP and LP requests. For HP requests, the distributions were close all along the distribution. For LP requests, there is some deviation as the percentage of calls taking less than 4 hours is slightly underestimated. Overall, both simulated distributions are close to the data. Numerically, the average response time for HP requests is 15.21 minutes in the data and 16.66 minutes in the simulation, for an error of 7.57%. For LP requests, the average response time is of 111.78 minutes in the data an of 120.26 minutes in the simulation, for an error of 10.05%. Both distributions were presented to US experts who deemed the model valid.

4.3.4 Estimation of the remaining transfer time

To implement the third policy, which includes "near-to-be-available" ambulances, we need to estimate the remaining transfer time (RTT) at the hospital. Two distinct considerations need to be addressed. First, how to estimate the average transfer time at the hospital. Second, how to use this estimate, and the information we have on how much time has been spent at the hospital thus far, to evaluate the RTT. It is worth noting that the latter must be evaluated each time an ambulance is a candidate for pre-assignment to a request and is thus a dynamic estimation.

The simplest method to estimate the average transfer time is to use the sample average over all requests, and all hospitals. However, this estimation can be improved by considering the information available for each transfer. Thus, we propose using a linear model with five categorical variables and one discrete variable to better predict the average transfert time at the hospital. The categorical variables are the hospitals, the priority of the requests, the day of the week, the hour of the day and a variable for the advancement of the paramedics' shift. This last variable was introduced after it was found that the time required to complete a transfer changed according to the time that remains for the paramedics' shift. The discrete variable is the number of ambulances that arrived up to 20 minutes before the current ambulance arrived, and represents the "queue" of ambulances. The regression provided a R^2 of 0.20 and every factor was found to be significant. The low value for R^2 was expected since the regression does not include variables such as the number of patients in the emergency department (ED) or their staffing level, which one could measure how busy the ED is at the time of the ambulance arrival.

Once the average $\overline{\mu}$ is found using one of the two methods, we must estimate the remaining transfer time RTT given that the ambulance has already spent t at the hospital. Again, two different methods are used. The first simply uses the difference between the average transfer time $\overline{\mu}$ and the time spents at the hospital thus far t. It should be noted that for a symmetrical

distribution, this estimation method will return zero in 50% of cases.

$$RTT = max(\overline{\mu} - t, 0) \tag{4.1}$$

For the second method, we consider the distribution of the transfer time. In our case, we found that it can be estimated with a normal distribution with an average of 52.3 minutes and a standard deviation of 13.6. Based on this information, the estimated remaining transfer time is based on the truncated average of the distribution for the time spent at the hospital :

$$RTT = E(X|X > t) - t \tag{4.2}$$

For a normal distribution, the truncated average can be expressed by this equation [34] :

$$\begin{aligned} \alpha = \frac{a - \overline{\mu}}{\overline{\sigma}}; \beta = \frac{b - \overline{\mu}}{\overline{\sigma}} \\ E(X|X > t) = \overline{\mu} - \overline{\sigma} \cdot \frac{\phi(0,1;\beta) - \phi(0,1;\alpha)}{\Phi(0,1;\beta) - \Phi(0,1;\alpha)} \end{aligned} \tag{4.3}$$

In this equation, the truncation range, the range for which we want to evaluate the average, is (α,β), the average of the non-truncated general distribution is $\overline{\mu}$ and its variance is $\overline{\sigma}$. ϕ represents the probability density function and Φ the cumulative density function. In our context, since $a = t$, and $\beta = +\infty$, the RTT can be estimated by :

$$RTT = \overline{\mu} + \overline{\sigma} \cdot \frac{\phi(0,1;\frac{t-\overline{\mu}}{\overline{\sigma}})}{1 - \Phi(0,1;\frac{t-\overline{\mu}}{\overline{\sigma}})} - t \tag{4.4}$$

Figure 4.7 presents a comparison of these two approaches to estimate the RTT dynamically. While the first approach will return a value of 0 for any value over the average, the second approach always gives a non-zero value. By combining the methods to estimate the average transfer time and the remaining transfer time, we have four possible approaches. Results found with these four approaches are presented in section 4.4.

4.4 Results

This section first presents the results obtained with the four methods proposed to estimate the RTT when the PADT policy is considered. Then, it presents and analyzes the impact of the four dispatching policies we propose on responses time for both HP and LP requests. Each policy is compared to the policy currently in place at US, as described in Section 4.2.2.

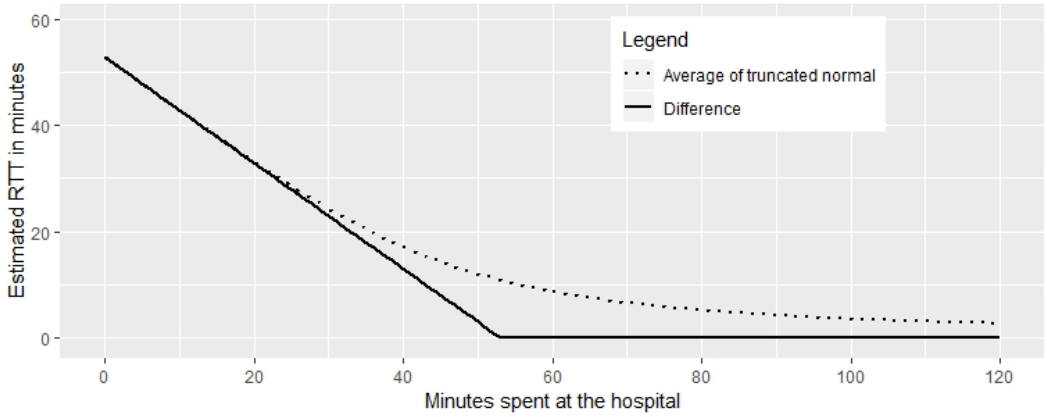

Figure 4.7 Comparison of estimation method for the remaining transfer time

This approach allows us to build upon state-of-the art dispatching policies.

Using the simulation model, we first compare the results obtained in terms of response times for both LP and HP requests, when the different methods are used to estimate the RTT for pre-assigment during transfers (PADT). We evaluated the impact of the PADT policy on HP RxT and LP RT using a series of t-tests. Each test compared the RT when using the PADT policy – using one of the methods to estimate RTT – to the current policy, also referred to as the reference scenario. As each methods, we use 60 replications of three months and include a 20-days warm-up period. Each scenario required a computing time of about 2 hours. Figures 4.8 and 4.9 present the results of the t-tests with a 95% confidence interval.

For HP RT, no significant effect was found with the given confidence interval, although the method using the regression to evaluate the mean and the difference to estimate the remaining transfer time is very close to a significant effect. For LP RT, all methods lead to a significant decrease, ranging from 6 to 9 minutes. The more complex method, using both the regression model to estimate the average, and the truncated average to estimate the RTT, gives the best result. Using the regression and difference provide a significative improvement over the use of the sample mean and the difference. As the method using both the regression and the truncated mean gave the best results, we selected it to compare the PADT to the other policies. Using the simulation model, we measured the effect of the four proposed policies on the response time when compared to reference scenario. Since the four policies could interact with each other, this leads to 16 combinations including the base scenario. For each scenario, we tested 60 replications of three months. A 20-day warm-up period was included in each replication. Table 4.1 presents the average effect on HP and LP RT, in minutes, for each factor combination. A '+' sign in the factor's column indicates that the policy was used in

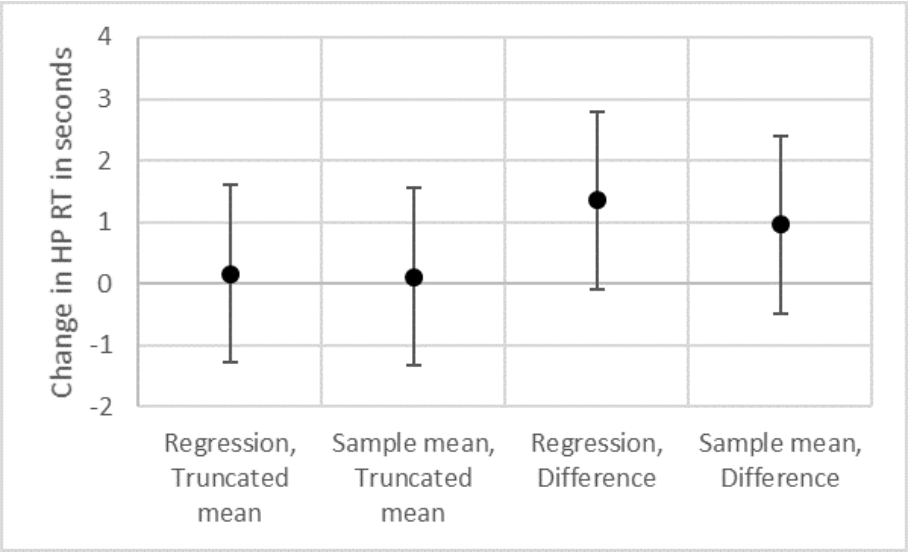

Figure 4.8 Effects of the estimation method on HP RT

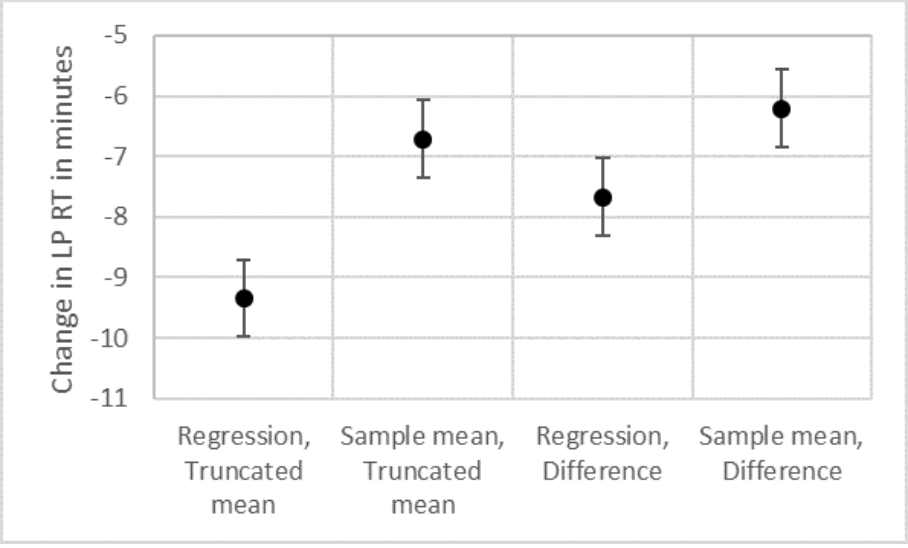

Figure 4.9 Effects of the estimation method on LP RT

Table 4.1 Results by factor combination

Factor Combination	e1 (PABS)	e2 (PADB)	e3 (PADT)	e4 (FAELP)	HP RT	LP RT
1	-	-	-	-	16.75	128.35
2	-	-	-	+	16.78	119.09
3	-	-	+	-	16.75	119.02
4	-	-	+	+	16.80	117.87
5	-	+	-	-	16.74	126.94
6	-	+	-	+	16.80	121.40
7	-	+	+	-	16.75	122.48
8	-	+	+	+	16.80	118.22
9	+	-	-	-	16.72	122.87
10	+	-	-	+	16.80	119.88
11	+	-	+	-	16.76	120.06
12	+	-	+	+	16.82	116.78
13	+	+	-	-	16.74	123.21
14	+	+	-	+	16.81	120.59
15	+	+	+	-	16.75	119.90
16	+	+	+	+	16.81	117.67

the combination while, a '-' sign indicates that it was not. Combination 1 corresponds to the reference scenario. Compared to this combination, many configurations provide an important decrease to RT for LP with negligible to no impact on response time for HP. Combination 16, which uses all proposed policies, provides a maximum reduction of 10.5 minutes, and scenario 4, which uses the PADT and FAELP policies, yields a similar result.

To better understand those results, we used the full 2^k factorial design method as proposed by Law [9]. For each replication of every scenario, the primary and secondary effects were measured. This created 60 different values for each effect. We averaged those values and estimated the error on this average with a 95% confidence interval. Figure 4.10 presents the results for HP RT, while 4.11 presents results for LP RT. The effects (e1 to e4) show the average changes in RT when their respective factor change from - to +. The two-factor interaction effect (e1e2 to e3e4) measures the impact of having the two factors in the same direction. For HP RT, the significative effects are e3, which causes an increase of up to 2 seconds, and e4 (FAELP), which causes a rise in RT of about 5 to 8 seconds. For LP RT, all scenarios reduce response time except e2 which increases it. Two-factor interaction is significative for e1e3, e1e4 and 3e4, as they cause an increase to RT, while e1, e3 and e4 reduce it. We can interpret the interaction as a reduction in the efficiency of each policy when more than one is used.

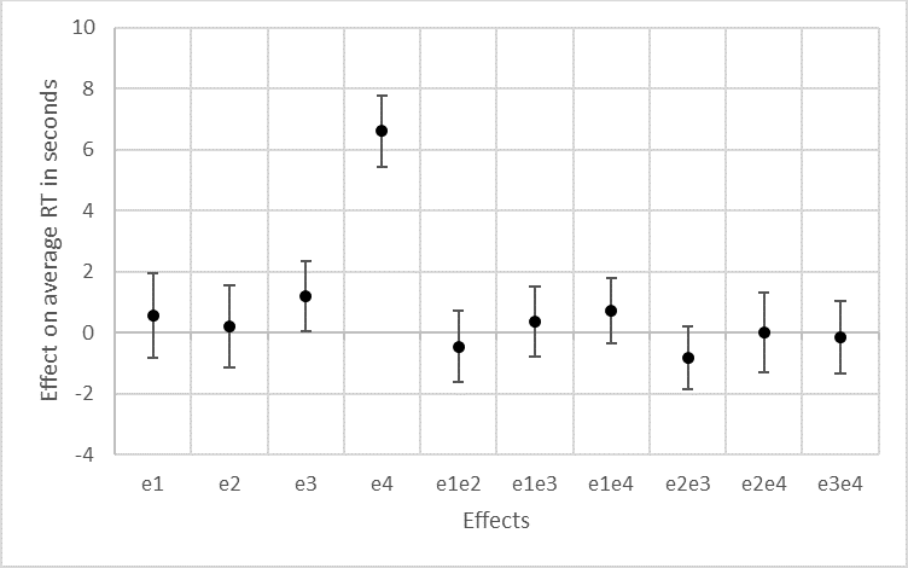

Figure 4.10 Effects of policies on HP RT in seconds

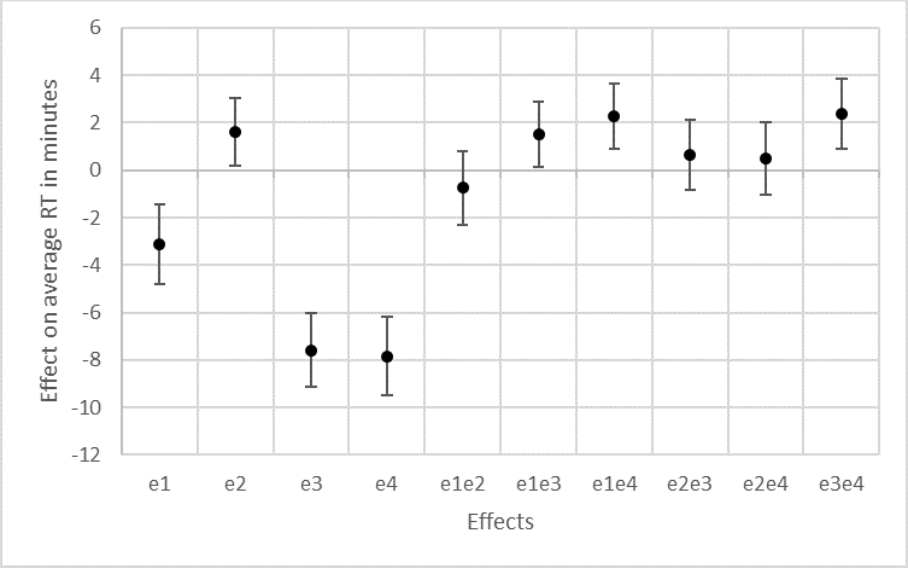

Figure 4.11 Effects of policies on LP RT in minutes

4.5 Discussions and managerial insights

In this paper, we proposed four policies to improve dispatching for EMS. The experiments we conducted show the positive impact of three of these policies, implemented alone or in parallel. The challenge in practice is determining which, when and how each of these policies should be selected. Recall that these policies aim to reduce the inefficiency that occurs each time an ambulance is assigned to a request that an unavailable ambulance would be able to reach sooner. When this happens, the system performance is impacted in two ways. First, the patient will have to wait longer before an ambulance reaches him, increasing RT. Secondly, since the travel time to the patient is longer than it would be if the unavailable ambulance were utilized, the overall time required to answer calls is increased, increasing the ambulance utilization rate and decreasing the ambulance availability.

The proposed policies could be implemented in a real-life setting assuming some conditions are met. First, acknowledging the diversity of the EMS environment, and uniqueness of each system, it is important to evaluate the relevance of these policies to a particular EMS. Furthermore, we implemented these policies in a system that is already efficient. To improve the performance of an EMS using more rudimentary dispatching policies, implementing some of the policies already in use by US first might be more impactful. Notably, the regular FAE policy should be implemented before considering our FAELP proposition. The smart-assignment policy should also be prioritized as it provides a very clear improvement to every priority [13]. Re-routing and coverage requirements also appear to be important as they help avoid a rise in HP RT as we incorporate our policies.

For an EMS already using FAE, implementing the FAELP would likely be quite simple ; however, it should be done carefully as we did measure a slight increase to HP RT with this policy. The PABS and PADB policies might be more complex as they require the dispatcher to consider busy ambulances, and as a result, some changes to the dispatching software are likely to be required. For the PADT policy, the exact complexity will depend on the estimation method, but as noted earlier even the most simple method provides a positive improvement. While implementing more than one policy would require more time, it is likely that some of the work used to implement one of the pre-assignment policies could be re-used to implement the other pre-assignment policies, as they are similar.

The three pre-assignment policies, while effective, have some limitations. The first is that not all unavailable ambulances are considered. For a variety of reasons, about 8% of emergency calls do not result in a transfer to a hospital. For those, after spending some time at the location of the call, the paramedic team will become available again. Pre-assignment is not

applicable to those ambulances since whether or not a transfer will be used is unpredictable. Furthermore, the RTT for the PADT policy is based on an estimate. This can result in cases where the pre-assigned ambulance ends up taking longer to answer the request than other ambulances could have. This is mitigated by the fact that whenever an ambulance becomes available, the system will check to see if the ambulance can reach any pre-assigned requests faster than the pre-assigned ambulance.

While pre-assignment aims to avoid the aforementioned inefficiency by anticipating ambulance availability, the free-ambulance dispatching policy is used to correct it. With FADLP, a newly available ambulance can be sent to an assigned patient provided it can reach the patient faster. In the cases without pre-assignment, this means that if an ambulance is close to the patient and about to become available, even though another ambulance was initially sent, the first ambulance might end up responding to the request. Compared to pre-assignment, free-ambulance exploitation does not completely avoid inefficiency, since the ambulance sent before the use of the policy will have lost some time traveling to the requests.

Figure 4.12 illustrates the behavior of both policies. A1 and A2 are both ambulances. A1 is initially busy for the next 2 minutes while A2 is idle. There are currently two patients in the system, P1 and P2, both are low priority. P1 is ahead of P2 in the requests queue and must be assisted first. If A1 is pre-assigned to P1, A2 will be dispatched to P2 which will result in the best outcome. Otherwise, A2 will become assigned to P1 while P2 remains unassigned. If free-ambulance exploitation is authorized, once A1 becomes available it will be sent to P1, which will lead to A2 being assigned to P2 instead. This result is a worse outcome than if A2 was sent directly to P2 as in the pre-assignment case. Both cases give much better results than if neither policy is used.

While the performance is clear in this example, those kinds of events will not necessarily occur frequently for all EMS. In a very small territory with short travel times, we might expect that idle ambulances could almost always reach the requests faster than busy ones. Also, in a system with a low utilization rate and a lot of available ambulances, those events would be uncommon. The first step for an EMS considering incorporation of any new policies would be to evaluate the frequency of these types of events.

The results show that PABS and PADT policy have a comparable effect, they reduce the RT of LP requests without affecting the HP RT. The PADB does not appear to improve RT. This is explained by the choice we made in the simulation model of where to position ambulances during the paramedics' break. They are often times at the same location as the other ambulances. This limits how often they can reach a request faster than the available ambulance. By contrast, the positive impact of the PABS and PADT policy can be explained

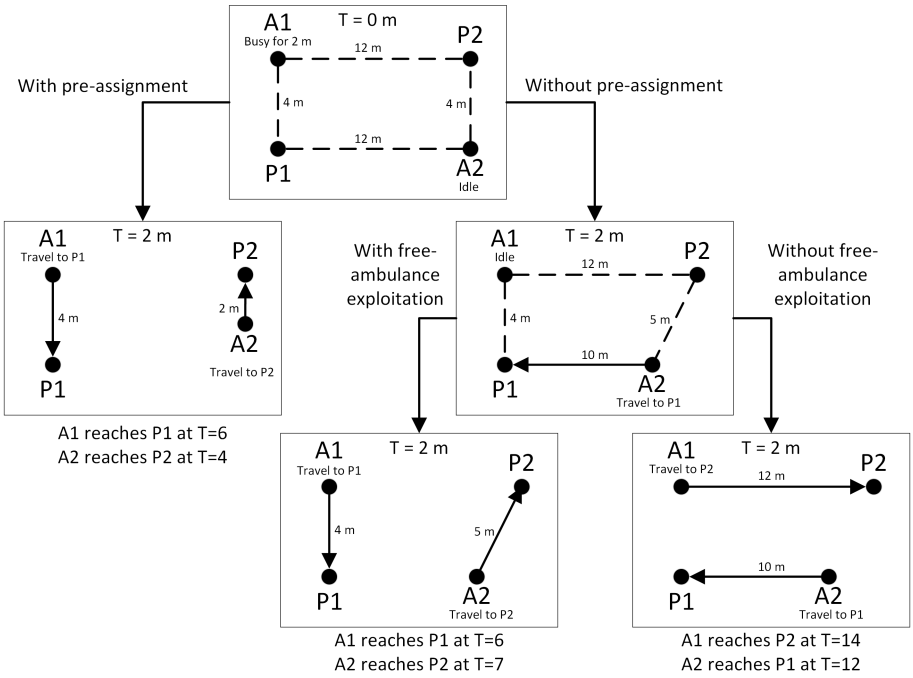

Figure 4.12 Example of how both pre-assigment and free-ambulance exploitation can improve RT and avoid unecessary travel

by the location of ambulances pre-assigned with those policies. The PABS policy uses ambulances that are about to start their shift. Three locations are possible for the shift start and none of them correspond to a waiting station. For the PADT, ambulances located at any of the 18 hospitals in the territory can be used. Again, those locations are not redundant with the waiting station. We may argue that if the lunch breaks were taken at a location other than the waiting station it's possible that the PADB would perform better.

Some positive two-factor interactions can be measured between the effect for a 95% confidence interval. This is normal since every policy aims to reduce the same inefficiency. Since implementing more than one policy would require more effort for an EMS, this might mean that it is preferable to not use too many of them at once. Notably, using the PADT and FAELP policies give nearly the same performance increase to LP RT than using all four policies.

The FAELP policy has the most important effect on LP RT, however, it is the only one that increases HP RT significantly. The increase is estimated to be around 5 to 8 seconds. This could be explained by the reduction in the travel time of ambulances traveling toward LP requests caused by this policy. Since re-routing is enabled, those ambulances could be used for HP dispatching during this travel time. Indirectly, this can lead to a small reduction

of availability for HP requests and therefore a small increase to RT. Whether or not this increase is justifiable would depend on the EMS's objectives.

For the PADT policy, we worked under the assumption that the emergency operations center has no information about the status of the paramedics during a transfer. This is currently the case for *Urgences-santé*. However, we suggest that it might be possible to ask the paramedics to communicate with the center at some specific point during the transfer. For example, once the patient has been transferred and the paramedics are getting ready for the next assignment. This could help improve the estimation of the time remaining at the hospital and therefore the policy performance. Regarding this estimation, the method using both the regression and the truncated average gave the best result. However, it might be challenging for an EMS to use this exact approach as it could be hard to implement in their dispatching software. Using instead either the "regression and difference" or the "sample average and truncated average" method still gives good results and even the most simple method using "sample average and difference" had a positive impact in our simulation.

4.6 Conclusion

The present article proposes four new policies for dispatching. Using a discrete-event simulation model, we measured their impact when added to state-of-the-art dispatching policies used by a large EMS. The model was build using data provided by the EMS and aim to reproduce their process accurately. We found that three of those policies would result in improvement for low priority response time ranging from 4.2 to 7.3%. Using more than one policy could reduce them by up to 8.32%. These policy had very little impact on hight priority response, the maximun increase observed was of 0.18%. For one of these policies, we also developed and compared different methods to evaluate the remaining transfer time. A method using a regression model to estimate the transfer time and using the average of a truncated distribution gave the best result although more simple methods were also effective. Regarding the implementation of those policies, we presented some practical considerations to take into account by EMS manager. For future work, a possible extension of the policy using pre-assignment during transfers policy would be to allow pre-assignment while an ambulance is traveling toward a hospital with a patient, this might be relevant for EMS with short transfer times. In cases where the exact time of the next availability of ambulance is known, pre-assignment could also be tested for high priority requests.

CHAPITRE 5 DISCUSSION GÉNÉRALE

Comme l'article propose de nouvelles politiques d'affectation, la description du modèle de simulation se concentre sur la gestion des affectations. Pour bien fonctionner, le modèle incorpore plusieurs autres politiques, dont plusieurs présentent également un intérêt de recherche. Cette section vise à présenter brièvement les politiques de localisation et de relocalisation, de sélection du centre hospitalier et de gestion des ressources. D'abord, nous brosserons un portrait sommaire de chaque politique et de leur fonctionnement à Urgences-santé. Ensuite, nous expliquerons leur implantation dans le modèle et les hypothèses utilisées. Finalement, nous discuterons de la littérature scientifique touchant à ces politiques et du potentiel de recherche de notre simulateur en lien avec ces politiques.

5.1 Localisation et relocalisation

Les politiques de localisation et de relocalisation déterminent la gestion des postes d'attente et de la couverture territoriale d'un SPU. La localisation correspond au fait de décider de l'emplacement des postes d'attente. C'est une décision à moyen terme, car l'emplacement des postes d'attente est normalement fixe. La relocalisation correspond au fait de choisir un poste d'attente auquel envoyer les ambulances lorsqu'elles redeviennent disponibles. C'est une décision qui doit être prise en temps réel.

Urgences-santé a positionné 15 postes d'attente divisés en 3 paliers de priorité sur son territoire. Le premier palier comporte 7 postes d'attente qui doivent être occupés en priorité. Si ces 7 postes sont occupés, les 5 postes du deuxième palier seront occupés par les prochaines ambulances disponibles. De même, si les deux premiers paliers sont comblés, les prochaines relocations se feront vers les 3 postes du dernier palier. Dans le cas où tous les postes sont occupés, bien que cela arrive rarement en pratique, une deuxième ambulance peut être envoyée à un poste d'attente. Les ambulances qui sont stationnées à un poste vont généralement y rester jusqu'à leur prochaine affectation, et il n'y a normalement pas de déplacement d'un poste d'attente à un autre. Cela peut mener à des situations où un des postes d'attente du premier palier est vide alors que certains de ceux des paliers moins prioritaires sont occupés.

Lorsqu'une ambulance redevient disponible et qu'aucune affectation n'est possible, une décision de relocalisation est prise. Normalement, une ambulance est envoyée au poste d'attente le plus proche du palier prioritaire dont tous les postes ne sont pas comblés. Cela permet aux paliers de se remplir dans l'ordre prévu en limitant les distances parcourues par les am-

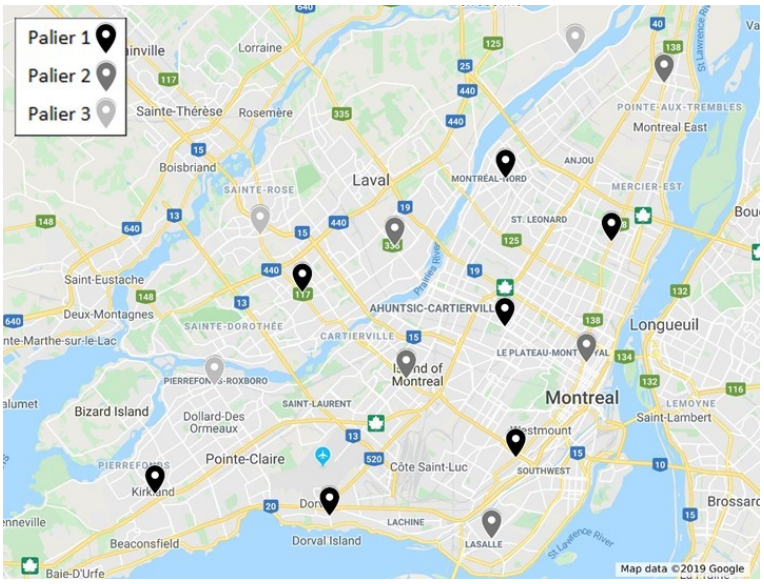

Figure 5.1 Localisation des postes d'attente

bulances. Une exception se produit lorsqu'une ambulance est dans les 45 dernières minutes de son quart : elle est alors envoyée à un poste d'attente qui la rapproche de son centre opérationnel, car elle devra bientôt s'y rendre.

L'occupation de ces postes d'attentes est en lien avec la politique de couverture territoriale d'US. Cette politique consiste à ne pas affecter d'ambulance à des appels de basse priorité tant qu'au moins 7 ambulances ne sont pas affectées à des postes d'attente. Ce critère ne tient pas compte des postes spécifiques occupés. Tant que le critère de couverture territoriale n'est pas satisfait, les ambulances qui redeviennent disponibles sont envoyées à des postes d'attente, à moins que des appels de hautes priorités ne soient en attente. Il est à noter qu'une ambulance n'a pas à être stationnée au poste pour être comptabilisée dans la couverture territoriale : elle peut également être en déplacement vers un poste d'attente.

Le modèle de simulation suit l'ensemble de la politique décrit ci-haut. Le modèle est donc fidèle aux politiques de relocalisation documentées par Urgences-santé. Il peut cependant exister certaines différences entre les politiques écrites et celles appliquées par les RMU. Par exemple, le seul poste d'attente du palier 1 situé à Laval est parfois priorisé par rapport aux autres postes du premier palier, car il assure une couverture minimale du territoire de Laval si aucune autre ambulance n'y est présente. Dû au manque de standardisation de ces pratiques, elles n'ont pas été intégrées au modèle.

La littérature met l'accent sur les politiques de localisation et relocalisation. Sur 29 articles de simulation mentionnés dans une revue de littérature [8], 21 touchent à un de ces deux

problèmes. Cependant, dans le cas d'US, ces politiques ne semblent pas être au cœur des préoccupations des gestionnaires. Cela peut s'expliquer par le taux d'utilisation des ambulances qui avoisine les 80%. Parce que les ambulances sont occupées la plupart du temps et qu'elles sont souvent immédiatement réaffectées dès qu'elles redeviennent disponibles, la gestion des ambulances disponibles est moins importante que dans d'autres SPU. En ce qui a trait aux études de localisation, la position des postes d'attente est souvent décidée à l'aide de méthodes d'optimisation, pour être ensuite testée dans un modèle de simulation [32, 35, 36].

Notre modèle pourrait donc être utilisé pour tester de nouvelles politiques de localisation, mais un travail important serait d'abord nécessaire pour définir ces politiques. Comme les décisions de relocalisation sont prises en temps réel, elles peuvent être définies et testées directement dans le simulateur sans passer par une étape d'optimisation hors du modèle. Puisqu'elle a été moins étudiée, la notion de couverture territoriale présente un certain intérêt de recherche. Présentement, la valeur de cette couverture est fixe dans le temps. Il pourrait être intéressant de varier sa valeur selon l'heure de la journée, car les conditions routières et le nombre de ressources disponibles fluctuent selon l'heure. Il semble raisonnable de supposer que la valeur optimale de ce critère change selon l'heure.

5.2 Sélection du centre hospitalier

La sélection du centre hospitalier est une décision qui est prise une fois les paramédics arrivés sur les lieux de l'appel. Le territoire d'US inclut un grand nombre d'hôpitaux dont plusieurs sont spécialisés. Normalement, les paramédics communiquent avec le CCS, et un RMU décide alors du centre hospitalier de destination à partir d'informations fournies par les paramédics et de critères prédéfinis. Il existe trois grandes catégories de cas : les cas spécifiques, les cas d'appartenance et les cas généraux.

Les patients classés dans la catégorie des cas spécifiques ont des besoins particuliers qui nécessitent leur envoi à un hôpital disposant d'une expertise particulière. Un certain nombre de cas spécifiques sont définis par Urgences-santé. Par exemple, les accidents vasculaires cérébraux, les amputations et les brûlures graves sont des cas spécifiques. Un ou plusieurs hôpitaux sur le territoire sont désignés pour chaque cas spécifique. Si un seul hôpital est désigné pour un cas spécifique, les patients y sont toujours envoyés. Si plusieurs hôpitaux sont éligibles, le plus proche est désigné dans la plupart des cas.

Les patients qui ne correspondent à aucun cas spécifique, mais qui sont suivis dans un centre hospitalier spécifique, sont classés comme des cas d'appartenance. Ces patients sont envoyés au centre hospitalier où ils sont suivis. Cette approche a pour but d'améliorer les soins au

patient dans l'hôpital même, car il est plus facile de prodiguer de meilleurs soins au patient si son dossier médical est accessible et s'il est connu du personnel et des médecins. Dans certains cas, les centres hospitaliers d'appartenance sont situés hors du territoire d'Urgences-santé, ce qui amène les paramédics à sortir du territoire d'US le temps de transporter le patient à son centre hospitalier d'appartenance et d'en revenir.

Les patients qui ne correspondent à aucun cas spécifique et à aucun cas d'appartenance sont classés comme des cas généraux. Ces patients peuvent être traités par n'importe quel centre hospitalier sur le territoire. US utilise ces cas pour équilibrer les volumes de patients envoyés à chaque hôpital du territoire. En effet, comme US est responsable de répartir les appels entre les différents centres hospitaliers, leur politique vise à répartir ces appels le plus équitablement possible afin d'éviter de surcharger les hôpitaux. Comme les règles pour les cas spécifiques et les cas d'appartenance ne laissent aucune marge de manœuvre dans la sélection de l'hôpital, les cas généraux servent à atteindre cet équilibre. Deux mécanismes sont en place à cet effet : les protections partielles et les quotes-parts. Chaque hôpital possède un nombre maximal de patients qui peut lui être envoyé en moins d'une heure. Ce nombre est défini dans une entente entre US et les centres hospitaliers de Montréal et de Laval. Si ce nombre est atteint, l'hôpital débute une période de protection partielle. Pendant cette période, aucun cas général ne peut lui être envoyé. Les quotes-parts, de leur côté, correspondent à un pourcentage maximal des appels qui peut être envoyé à un hôpital depuis les 24 dernières heures. Si un hôpital est au-dessus de sa quote-part ou, en d'autres mots, s'il a reçu un plus grand pourcentage des patients que sa quote-part, il est moins à risque de recevoir des cas généraux. US sélectionne donc un centre hospitalier pour chaque patient en se basant sur les protections partielles, les dépassements de quote-part et l'emplacement des patients par rapport aux hôpitaux. En plus des protections partielles, il arrive également qu'un hôpital demande à ne recevoir aucun patient pour une certaine période de temps, indépendamment du type de priorité de l'appel. Cette pratique est désignée sous le nom de protection complète.

Dans le modèle, les trois types de cas sont simulés. Pour les cas spécifiques et les cas d'appartenance, les listes des destinations possibles pour chaque type de cas sont définies conformément aux politiques d'US, et les ambulances sont envoyées à l'hôpital le plus proche parmi cette liste. Pour les cas généraux, les quotes-parts et les protections partielles sont modélisées en utilisant les seuils définis dans l'entente entre US et les hôpitaux du territoire. Les protections complètes ne sont cependant pas intégrées dans le modèle à cause de difficultés avec les données sur le sujet.

Le terme « protection complète », utilisé par US pour désigner l'action d'un hôpital de refuser des patients, est connu, dans la littérature, sous le nom de *Ambulance diversion* ou de

déroutement des ambulances. Plusieurs chercheurs se sont penchés sur cette problématique. Le déroutement peut être vu comme une solution au débordement des urgences des hôpitaux [37,38]. Il met en quelque sorte en opposition les besoins des centres hospitaliers, soit d'éviter les débordements, et ceux des SPU, soit de se rendre au centre hospitalier le plus près pour réduire les temps de transport [39]. Il a également été suggéré que le déroutement peut nuire aux patients, car cela peut retarder l'heure à laquelle ils reçoivent des soins [40]. Ramirez-Nafarrate [41] propose une approche qui concilie ces besoins divergents en proposant des politiques de déroutement qui visent à balancer le temps de transport et le temps d'attente à l'urgence des patients dans le but de minimiser les temps d'attente totaux.

Comme les protections complètes ne sont pas intégrées au simulateur, il serait nécessaire de les ajouter avant de contribuer à l'étude des politiques de déroutement. Une fois cela fait, le modèle permettrait de s'intéresser à ces politiques dans un contexte très réaliste et détaillé, comme dans celui d'une étude de cas. Urgences-santé possède une politique de sélection du centre hospitalier qui est, à notre connaissance, unique, en ce sens qu'elle distingue les cas spécifiques, les cas d'appartenance et les cas généraux. L'étude de ces politiques pourrait constituer une contribution intéressante à la recherche. En particulier, considérer les cas spécifiques et les cas d'appartenance dans la prise de décisions optimales pour les patients, tout en utilisant les cas généraux pour balancer le volume des appels envoyés aux hôpitaux, mérite d'être étudié plus en détails.

5.3 Gestions des effectifs

Les équipes de paramédics débutent leur quart de travail dans un des trois centres opérationnels d'Urgences-santé. Dès le début de leur quart, ils sont considérés comme disponibles. Ils sont alors gérés de la même manière qu'une ambulance de nouveau disponible après une affectation tel que présenté dans la section 4.2. Les quarts de travail ont normalement une durée de 8, 10 ou 12 heures, auquel peut s'additionner du temps supplémentaire.

Pendant le quart de travail, la plupart des équipes ont droit à une pause repas. Comme il est difficile de prévoir la disponibilité des paramédics, la pause n'est pas prise à une heure fixe. Les quarts de travail qui incluent une période de repas ont une « fenêtre d'opportunité », pendant laquelle la pause peut être prise. À partir du début de cette période, la période de repas peut débuter si les équipes sont disponibles et qu'aucun appel de haute priorité ne nécessite leur attention. Auquel cas, les équipes peuvent débuter une « période de rapprochement » de 15 minutes qui mène à leur pause repas. Pendant cette période, les paramédics peuvent déplacer leur ambulance librement vers l'endroit où ils souhaitent prendre leur repas. Pendant la période de rapprochement, les équipes peuvent être affectées aux appels les

plus urgents si aucune autre équipe ne peut répondre à l'appel dans des délais acceptables. Dans ces cas, la période de rapprochement est interrompue et sera reprise dans son entièreté par la suite. Les paramédics peuvent également mettre fin à la période de rapprochement en tout temps pour débuter immédiatement leur pause repas, auquel cas le temps restant à cette période est perdu. Pour les paramédics, cette option est intéressante, car elle permet d'éviter les interruptions du rapprochement. Une fois la période de rapprochement terminée, la vraie pause repas débute. Elle peut durer 30, 45 ou 60 minutes selon le quart de travail. Pendant cette période, l'ambulance ne peut être affectée à aucun appel. La seule possibilité d'interruption survient si les paramédics sont témoins d'une situation d'urgence ou si une personne leur demande de l'aide. Si les paramédics n'ont pas encore pu prendre leur pause repas après un certain point durant le période d'opportunité, le critère pour débuter la période de rapprochement change légèrement. La pause repas peut alors débuter même lorsque certains appels de haute priorité moins urgent sont pas en attente. Cet ajout aide les ambulanciers à prendre leur pause repas pendant une période d'achalandage. Les pauses repas sont considérées comme ayant été prises à l'heure si elles ont débuté dans la fenêtre d'opportunité mentionnée plus haut, et en retard si elle débute après la fin de cette fenêtre. Pendant des journées très achalandées, il peut arriver que les paramédics ne se fassent jamais accorder leur pause, auquel cas la pause repas est payée en temps supplémentaire.

En dehors des pauses repas, le quart de travail se déroule normalement jusqu'à ce qu'il reste moins de 45 minutes au quart de travail. Une fois cette limite passée, l'ambulance est considérée comme en fin de quart de travail. Dès ce moment, l'ambulance n'est plus disponible pour répondre à certaines priorités. La disponibilité est réduite davantage à chaque 15 minutes subséquentes. Lorsque le quart est terminé, l'ambulance n'est plus disponible. Ces restrictions ont pour but d'éviter les affectations d'ambulance qui terminent leur quart de travail bientôt, car cela entraîne du temps supplémentaire. Pendant cette période, les RMU vont également rapprocher les ambulances disponibles de leur centre opérationnel afin qu'elle puisse y terminer leur quart à l'heure. Les ambulances terminent toujours dans le centre opérationnel où elles ont commencé leur quart de travail.

Les deux paragraphes précédents décrivent les politiques de gestion des effectifs en situation normale. Il arrive parfois que des mesures spéciales soient prises lorsque le SPU est surchargé. Ces mesures peuvent inclure l'annulation de pauses repas ou du temps supplémentaire obligatoire.

Les pauses repas et les périodes de rapprochement qui les précèdent sont considérées dans le simulateur. Bien que les périodes de rapprochement puissent être raccourcies à la demande des paramédics, elles sont fixées à 15 minutes dans le simulateur. Nous avons fait cette

hypothèse, car nous n'avions pas de statistique sur leurs durées réelles, sans compter que plusieurs employés d'US nous ont affirmé qu'en pratique les paramédics demandent rarement que leur rapprochement soit raccourci. De la même manière, à cause d'un manque de données à ce sujet, nous n'avons pas modélisé les déplacements de l'ambulance pendant cette période. Les ambulances sont donc stationnées dans les postes d'attente pendant leur pause repas. Quant aux fins de quart, les politiques exactes d'US sont utilisées dans le simulateur.

Le simulateur ne considère pas directement les mesures spéciales. Cependant, comme nous reproduisons l'historique des quarts de travail, les conséquences de ces mesures sur les quarts de travail sont reproduites. Si, pendant une journée, des mesures spéciales ont limité les pauses repas, lorsque le simulateur reproduira cette journée, la même limitation aura lieu. Cette méthode nous apparaît acceptable, car ces mesures sont déployées lorsqu'il y a un débalancement entre le nombre d'appels et les véhicules disponibles. Comme ces deux facteurs suivent les valeurs historiques, ces mesures devraient se produire de manière cohérente avec leurs conditions d'enclenchement.

À notre connaissance, les politiques de gestion des pauses repas et des fins de quart des équipes de paramédics n'ont jamais été étudiées dans le contexte d'une simulation. Le fait de ne pas avoir de temps pour prendre une pause repas et celui de devoir faire du temps supplémentaire ont été liés à une augmentation du stress chez les paramédics dans d'autres SPU [42, 43]. Il pourrait donc être pertinent d'optimiser les politiques qui entourent la prise des repas et la fin de quart de manière à réduire ces deux problèmes. Une telle démarche soulèverait une problématique intéressante dans la mesure où le fait d'améliorer les conditions de travail risque d'avoir un impact sur le service à la population. Il faudrait donc parvenir à balancer ces deux nécessités dans une telle étude. De manière plus large, la planification des horaires des ambulances a été étudiée à quelques reprises, parfois en combinaison avec un problème de localisation [44–46]. Tout comme pour les problèmes de localisation, le problème de planification n'est pas une décision prise en temps réel. Elle peut donc être étudiée en dehors d'un modèle de simulation. La simulation peut cependant servir à tester la qualité des solutions développées.

5.4 Spécialisation des ressources

Comme plusieurs autres SPU, Urgences-santé possède une flotte qui comporte plusieurs types de ressources. D'abord, il faut distinguer les ambulances des autres ressources, puisque seules les ambulances peuvent transporter des patients. Les ambulances sont elles-mêmes divisées en trois catégories : les véhicules 911, les véhicules Inter et les véhicules Hybride. Les véhicules 911 ont pour priorité de répondre aux appels d'urgence transmis par le 911, alors que

les véhicules Inter sont spécialisés dans le transport interétablissement et ne répondent que minimalement aux appels d'urgence. Bien que leur spécialisation tire plus vers l'interétablissement, les véhicules Hybrides sont, de leur côté, un entre-deux entre les véhicules 911 et les véhicules Inter. Les véhicules de type 911 constituent 96% de la flotte ambulancière alors que les deux autres catégories représentent environ 2% chacune.

La spécialisation des ressources ambulancières est gérée dans les politiques d'affectation par des règles spécifiques à chaque priorité. Pour rappel, US utilise 9 priorités, dont 5 sont liées aux appels d'urgence et 4 aux transports interétablissements. Pour les priorités liées aux appels d'urgence, l'affectation de véhicules 911 est favorisée par rapport à l'affectation de véhicules Hybrides, tandis que les véhicules Inter ne peuvent simplement pas être affectés à ces appels dans la plupart des cas. Pour les demandes de transport interétablissement, les véhicules Hybrides et Inters sont privilégiés par rapport aux véhicules 911. Comme les demandes de transport interétablissement représentent environ 10% des demandes, alors que les véhicules Inters et Hybrides forment ensemble 4% de la flotte, les véhicules 911 répondent à plus de la moitié de ces demandes.

À part les ambulances, plusieurs autres types de ressources sont également utilisés. Notamment, les équipes de paramédics en soins avancés peuvent assister les équipes ambulancières sur certaines demandes. Ces équipes sont constituées de deux paramédics dont la formation leur permet de fournir des soins préhospitaliers avancés. Ces équipes s'affectent elles-mêmes à des demandes qui nécessitent leurs compétences. Comme les équipes ambulancières, ils se déplacent vers les lieux de l'incident et prodiguent des soins préhospitaliers. Un paramédic en soins avancés peut également monter dans une ambulance et continuer à prodiguer des soins pendant le transport à l'hôpital pendant que son coéquipier suit l'ambulance dans leur véhicule. US possède aussi une unité de soutien opérationnel dont le rôle est d'aider les paramédics à atteindre des patients donc l'accès est difficile ou de déplacer des patients souffrant d'obésité morbide. Contrairement aux paramédics de soins avancés, leurs interventions terminent une fois le début du transport par ambulance vers l'hôpital. Un groupe d'intervention médical tactique peut également être utilisé lors d'intervention à haut risque, souvent conjointement avec des équipes policières. Une équipe de constat de décès, dans laquelle travaille un médecin, peut également être appelée par des paramédics pour des constats de décès.

Le modèle de simulation inclut les trois types d'ambulances et gère leurs spécificités selon les politiques d'US. Les équipes non-ambulancières ne sont cependant pas modélisées directement, car elles n'effectuent pas de transport et sont normalement affectées conjointement à une équipe ambulancière. Leur effet sur les temps de réponse et sur les autres indicateurs mesurés dans le modèle est faible. Cependant, comme le temps passé sur les lieux des incidents

se base sur les données historiques, les délais, occasionnés par le fait d'attendre l'arrivée d'une équipe spécialisée avant de débuter le transport vers l'hôpital, sont inclus dans le modèle.

Dans la littérature, l'utilisation de deux classes de véhicule, une pour les soins réguliers (*Basic Life Support*) et une autre pour les soins avancés (*Advanced Life Support*), a déjà été étudiée [47–49]. L'utilisation des deux types de véhicule peut varier d'une étude à l'autre. Il est possible que seules les équipes de soins avancés puissent faire du transport [49], que les deux types puissent en faire [48] ou, comme c'est le cas à US, que seules les équipes de soins réguliers s'occupent du transport [47]. Il pourrait être intéressant de comparer ces différents modèles. Dans le cas d'US, les véhicules de soins avancés ne s'occupent pas du transport, ce qui leur permet de répondre à plus d'appels qui nécessitent leurs compétences. Malgré cet avantage, cette approche réduit le nombre de véhicules disponibles pour du transport, ce qui peut réduire la capacité du système à répondre à des pics de demande. Cette approche est causée en partie au petit nombre d'équipes de soins avancés disponibles. Cependant, il est possible que, si des équipes de soins avancés s'ajoutent, il devient plus efficace d'utiliser les véhicules de soins avancés pour du transport. Il pourrait être pertinent de comparer ces deux approches avec différents ratios de soins réguliers et avancés. De la même manière, l'utilisation de véhicules spécialisés pour les transferts interétablissements et les appels d'urgence a déjà été examinée [29]. L'article compare l'utilisation d'ambulances complètement spécialisés à des méthodes utilisant une flotte commune pour ces deux types de demandes. Les résultats montrent que la spécialisation complète des ressources est moins efficace en matière de qualité du service que les alternatives proposées. Dans le même ordre d'idées, il pourrait être pertinent de reconsidérer l'utilisation des véhicules spécialisés à US.

5.5 Sommaire de la discussion

Le modèle de simulation intègre plusieurs aspects en dehors des politiques d'affectation. Les politiques de localisation et de relocalisation, de sélection du centre hospitalier, de gestion des effectifs et de spécialisation des ressources sont considérées en détails. Le modèle reproduit au mieux le fonctionnement d'US d'après les informations fournies. Les hypothèses utilisées ont été validées auprès d'experts d'US. Il en résulte un modèle complexe et réaliste qui a le potentiel de contribuer à la recherche sur diverses politiques.

CHAPITRE 6 CONCLUSION ET RECOMMENDATIONS

Ce mémoire présente un modèle de simulation basé sur le SPU Urgences-santé. Le modèle est utilisé pour tester quatre nouvelles politiques d'affectation des ambulances. Les trois premières politiques consistent à pré-affecter les ambulances avant le début des quarts de travail, pendant les pauses repas et lors des transferts de patient. La dernière politique consiste à utiliser les ambulances nouvellement disponibles pour remplacer des ambulances déjà affectées lorsque cela amène des gains de temps. Ces politiques sont utilisées pour les appels de basses priorités. À l'exception de la pré-affectation d'ambulances pendant les pauses repas, toutes les politiques ont un impact positif sur les temps de réponse des appels de basses priorités tout en gardant les temps de réponse de hautes priorités stables.

Pour arriver à ces résultats, nous avons suivi les étapes standards d'un projet de simulation. Dans un premier temps, nous avons formulé notre problème, défini le modèle et ses hypothèses et récolté des données. L'ensemble des informations nécessaires à cet effet a été fourni par Urgences-santé. Nous avons également validé nos hypothèses et présenté notre définition du modèle avec US. Par la suite, nous avons développé et validé le modèle. La validation repose à la fois sur une analyse numérique et une présentation du modèle à des experts des processus d'Urgences-santé. Ensuite, nous avons défini les quatre nouvelles politiques qui font l'objet de l'article. Ces quatre politiques et leurs interactions ont été testées à l'aide du modèle. Enfin, les résultats de ces expérimentations ont été analysés et documentés.

Le modèle de simulation, bien que réaliste, repose sur plusieurs hypothèses. De manière générale, nous avons suivi la documentation d'Urgences-santé et les instructions données aux répartiteurs médicaux d'urgence pour reproduire leurs décisions. Cependant, il peut exister des écarts entre ces normes et la réalité. Certains éléments plus précis du fonctionnement d'Urgences-santé n'ont également pas été modélisés. Notamment, les déplacements pendant la période de rapprochement, qui précède le repas, ne sont pas inclus. Cette limitation a vraisemblablement un impact sur les résultats de la politique qui pré-affectent les ambulances pendant la période de repas. Inclure ces déplacements serait d'ailleurs une amélioration possible des présentes recherches.

Dans un autre ordre d'idées, le modèle de simulation pourrait être utilisé pour évaluer des politiques en dehors de l'affectation. La politique de couverture territoriale présente un certain intérêt et nous pourrions étudier la possibilité de la faire varier dans le temps. La politique de sélection des centres hospitaliers, qui cherche à distribuer équitablement les patients entre les hôpitaux, tout en envoyant les patients dans les centres les plus appropriés, pourrait

être étudiée plus en détails. Nous pourrions potentiellement mesurer son coût en termes de performance ou tenter d'améliorer leur méthode pour répartir les patients équitablement. Améliorer les politiques sur les pauses repas et les fins de quart présente également un intérêt étant donné qu'elles n'ont jamais été étudiées à notre connaissance. Finalement, l'étude des ressources spécialisées, tant en termes de soins de base et de soins avancés qu'en termes de ressources dédiées aux appels d'urgence et à l'interétablissement, est envisageable.

RÉFÉRENCES

[1] Urgences-santé, "Rapport annuel de gestion," Urgences-santé, Rapport technique, 2018. [En ligne]. Disponible : https://www.urgences-sante.qc.ca/wp-content/uploads/2018/12/Rapport-annuel-2017-2018-PDF.pdf

[2] A. Bürger *et al.*, "The effect of ambulance response time on survival following out-of-hospital cardiac arrest : An analysis from the german resuscitation registry," *Deutsches Ärzteblatt International*, vol. 115, n°. 22-24, p. 541–548, août 2018.

[3] A. A. Báez *et al.*, "Predictive effect of out-of-hospital time in outcomes of severely injured young adult and elderly patients," *Prehospital and disaster medicine*, vol. 21, n°. 6, p. 427–430, nov. 2006.

[4] E. T. Wilde, "Do emergency medical system response times matter for health outcomes ?" *Health economics*, vol. 22, n°. 7, p. 790–806, juill. 2013.

[5] "Faute d'ambulance, une dame âgée demeure au sol plus de 3 heures," *TVA Nouvelles*, 2017. [En ligne]. Disponible : https://www.tvanouvelles.ca/2017/02/19/exclusif--faute-dambulance-une-dame-agee-demeure-au-sol-plus-de-3-heures

[6] P. Lagacé, "10 heures d'attente, 10," *La Presse*, 2018. [En ligne]. Disponible : http://plus.lapresse.ca/screens/ccf25e00-3fc7-48a7-8bea-63dae6ac7f65___7C____0.html

[7] F. Giguère, "Elle attend 7 heures pour avoir une ambulance," *Le Journal de Montréal*, 2016. [En ligne]. Disponible : https://www.journaldemontreal.com/2016/02/10/elle-attend-7-heures-pour-avoir-une-ambulance

[8] L. Aboueljinane, E. Sahin et Z. Jemai, "A review on simulation models applied to emergency medical service operations," *Computers & Industrial Engineering*, vol. 66, n°. 4, p. 734–750, déc. 2013.

[9] A. M. Law, *Simulation modeling and analysis*, 5e éd. New York, USA : McGraw–Hill Education, 2015.

[10] V. Bélanger, A. Ruiz et P. Soriano, "Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles," *European Journal of Operational Research*, vol. 272, n°. 1, p. 1–23, janv. 2019.

[11] T. Andersson et P. Värbrand, "Decision support tools for ambulance dispatch and relocation," *Journal of the Operational Research Society*, vol. 58, n°. 2, p. 195–201, déc. 2007.

[12] S. Yoon et L. A. Albert, “An expected coverage model with a cutoff priority queue. health care management science,” *Health care management science*, vol. 21, n°. 4, p. 517–533, déc. 2018.

[13] R. Aringhieri *et al.*, “A simulation and online optimization approach for the real-time management of ambulances,” communication présentée à (WSC ’18) Proceedings of the 2018 Winter Simulation Conference, Gothenburg, Sweden, 9-12 dec 2018, p. 2554–2565. [En ligne]. Disponible : https://ieeexplore.ieee.org/document/8632231

[14] C. S. Lim, R. Mamat et T. Braunl, “Impact of ambulance dispatch policies on performance of emergency medical services,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, n°. 2, p. 624–632, juin 2011.

[15] M. Gendreau, G. Laporte et F. Semet, “A dynamic model and parallel tabu search heuristic for real-time ambulance relocation,” *Parallel computing*, vol. 27, n°. 12, p. 1641–1653, nov. 2001.

[16] S. G. Henderson et A. J. Mason, “Ambulance service planning : simulation and data visualisation,” dans *Operations Research and Health Care*, M. L. Brandeau, F. Sainfort et W. P. Pierskalla, édit. Boston, USA : Springer, 2005, p. 77–102. [En ligne]. Disponible : http://dx.doi.org/10.1007/1-4020-8066-2_4

[17] K. Shin, I. Sung et T. Lee, “Emergency medical service system design evaluator,” communication présentée à (WSC ’13) Proceedings of the 2013 Winter Simulation Conference : Simulation : Making Decisions in a Complex World, Washington, D.C., USA, 8-11 dec 2013, p. 2410–2421. [En ligne]. Disponible : https://ieeexplore.ieee.org/document/6721615

[18] S. Lee, “Role of parallelism in ambulance dispatching,” *IEEE Transactions on Systems, Man, and Cybernetics : Systems*, vol. 44, n°. 8, p. 1113–1122, janv. 2014.

[19] N. B. J. M. Theeuwes *et al.*, “Formalization and improvement of ambulance dispatching in brabant-zuidoost,” mémoire de maîtrise, Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, Eindhoven, Pays-Bas, 2019. [En ligne]. Disponible : https://research.tue.nl/en/studentTheses/formalization-and-improvement-of-ambulance-dispatching-in-brabant

[20] D. Bandara, M. E. Mayorga et L. A. McLay, “Priority dispatching strategies for ems systems,” *Journal of the Operational Research Society*, vol. 65, n°. 4, p. 572–587, avr. 2014.

[21] A. A. Nasrollahzadeh, A. Khademi et M. E. Mayorga, “Real-time ambulance dispatching and relocation,” *Manufacturing & Service Operations Management*, vol. 20, n°. 3, p. 467–480, avr. 2018.

[22] S. Lee, “The role of preparedness in ambulance dispatching,” *Journal of the Operational Research Society*, vol. 62, n°. 10, p. 1888–1897, oct. 2011.

[23] V. Schmid, “Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming,” *European journal of operational research*, vol. 219, n°. 3, p. 611–621, juin 2012.

[24] C. J. Jagtenberg, S. Bhulai et R. D. van der Mei, “Dynamic ambulance dispatching : is the closest-idle policy always optimal ?” *Health care management science*, vol. 20, n°. 4, p. 517–531, déc. 2017.

[25] I. E. Blanchard *et al.*, “Emergency medical services response time and mortality in an urban setting,” *Prehospital Emergency Care*, vol. 16, n°. 1, p. 142–151, janv. 2012.

[26] S. S. W. Lam *et al.*, “Factors affecting the ambulance response times of trauma incidents in singapore,” *Accident Analysis & Prevention*, vol. 82, p. 27–35, sept. 2015.

[27] A. Morse, “Transforming nhs ambulance services,” National Audit Office, Rapport technique HC 1086 Session 2010–2012, 2011. [En ligne]. Disponible : https://www.nao.org.uk/wp-content/uploads/2011/06/n10121086.pdf

[28] S. Enayati *et al.*, “Ambulance redeployment and dispatching under uncertainty with personnel workload limitation,” *International Transactions in Operational Research*, vol. 50, n°. 9, p. 777–788, 2018.

[29] Y. Kergosien *et al.*, “A generic and flexible simulation-based analysis tool for ems management,” *International Journal of Production Research*, vol. 53, n°. 24, p. 7299–7316, avr. 2015.

[30] L. Aboueljinane *et al.*, “A simulation study to improve the performance of an emergency medical service : application to the french val-de-marne department,” *Simulation modelling practice and theory*, vol. 47, p. 46–59, sept. 2014.

[31] S. S. W. Lam *et al.*, “Reducing ambulance response times using discrete event simulation,” *Prehospital Emergency Care*, vol. 18, n°. 2, p. 207–216, avr. 2014.

[32] R. Aringhieri, G. Carello et D. Morale, “Ambulance location through optimization and simulation : the case of milano urban area,” communication présentée à XXXVIII Annual Conference of the Italian Operations Research Society Optimization and Decision Sciences, 2007, p. 1–29. [En ligne]. Disponible : https://air.unimi.it/retrieve/handle/2434/40782/6875/aor.pdf

[33] Y. M. Carson et R. Batta, “Locating an ambulance on the amherst campus of the state university of new york at buffalo,” *Interfaces*, vol. 20, n°. 5, p. 43–49, oct. 1990.

[34] J. Burkardt, *The truncated normal distribution.* USA : Department of Scientific Computing Website, Florida State University, 2014. [En ligne]. Disponible : https://people.sc.fsu.edu/~jburkardt/presentations/truncated_normal.pdf

[35] S. Nickel, M. Reuter-Oppermann et F. Saldanha-da Gama, "Ambulance location under stochastic demand : A sampling approach," *Operations Research for Health Care*, vol. 8, p. 24–32, mars 2016.

[36] H. Leknes *et al.*, "Strategic ambulance location for heterogeneous regions," *European Journal of Operational Research*, vol. 260, n°. 1, p. 122–133, juill. 2017.

[37] C. H. Lin, C. Y. Kao et C. Y. Huang, "Managing emergency department overcrowding via ambulance diversion : A discrete event simulation model," *Journal of the Formosan Medical Association*, vol. 114, n°. 1, p. 64–71, janv. 2015.

[38] D. Pförringer *et al.*, "Closure simulation for reduction of emergency patient diversion : a discrete agent-based simulation approach to minimizing ambulance diversion," *European journal of medical research*, vol. 23, n°. 1, p. 1–8, juin 2018.

[39] E. Mund, "Ending ambulance diversion. eighteen hospitals in king county, wash., work toward a perpetual zero-divert status," *EMS world*, vol. 40, n°. 4, p. 31–38, avr. 2011.

[40] M. Li, P. Vanberkel et C. A. J. E., "A review on ambulance offload delay literature," *Health Care Management Science*, vol. 22, n°. 4, p. 658–675, juill. 2018.

[41] A. Ramirez-Nafarrate, J. W. Fowler et T. Wu, "Design of centralized ambulance diversion policies using simulation-optimization," communication présentée à (WSC '11) Proceedings of the 2011 Winter Simulation Conference, Phoenix, USA, 11-14 dec 2011, p. 1251–1262. [En ligne]. Disponible : https://ieeexplore.ieee.org/document/6147846

[42] T. Sterud *et al.*, "Occupational stressors and its organizational and individual correlates : a nationwide study of norwegian ambulance personnel," *BMC emergency medicine*, vol. 8, n°. 1, p. 1–11, janv. 2008.

[43] K. Mahony, "Restructuring and the production of occupational stressors in a corporatised ambulance service," *Health Sociology Review*, vol. 14, n°. 1, p. 84–96, juin 2005.

[44] H. K. Rajagopalan *et al.*, "Ambulance deployment and shift scheduling : An integrated approach," *Journal of Service Science and Management*, vol. 4, n°. 1, p. 66–78, mars 2011.

[45] R. McCormack et G. Coates, "A simulation model to enable the optimization of ambulance fleet allocation and base station location for increased patient survival," *European Journal of Operational Research*, vol. 247, n°. 1, p. 294–309, nov. 2015.

[46] J. L. Vile *et al.*, “Time-dependent stochastic methods for managing and scheduling emergency medical services,” *Operations Research for Health Care*, vol. 8, p. 42–52, mars 2016.

[47] C. O. A. Stein, “Emergency medical service response system performance in an urban south african setting : a computer simulation model,” thèse de doctorat, University of Cape Town, Rondebosch, Afrique du Sud, 2014. [En ligne]. Disponible : https://open.uct.ac.za/handle/11427/9523

[48] S. Su et C. L. Shih, “Modeling an emergency medical services system using computer simulation,” *International journal of medical informatics*, vol. 72, n°. 1-3, p. 57–72, déc. 2003.

[49] K. Sudtachat, M. E. Mayorga et L. A. McLay, “Recommendations for dispatching emergency vehicles under multitiered response via simulation,” *International Transactions in Operational Research*, vol. 21, n°. 4, p. 581–617, mars 2014.